



Information Systems Research

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Predicting Labor Market Competition: Leveraging Interfirm Network and Employee Skills

Yuanyang Liu, Gautam Pant, Olivia R. L. Sheng

To cite this article:

Yuanyang Liu, Gautam Pant, Olivia R. L. Sheng (2020) Predicting Labor Market Competition: Leveraging Interfirm Network and Employee Skills. Information Systems Research 31(4):1443-1466. <https://doi.org/10.1287/isre.2020.0954>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2020, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.



For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Predicting Labor Market Competition: Leveraging Interfirm Network and Employee Skills

Yuanyang Liu,^a Gautam Pant,^b Olivia R. L. Sheng^c

^a Department of Business Analytics and Statistics, Haslam College of Business, University of Tennessee, Knoxville, Tennessee 37996;

^b Department of Business Analytics, University of Iowa, Iowa City, Iowa 52242; ^c Department of Operations and Information Systems, University of Utah, Salt Lake City, Utah 84112

Contact: yliu191@utk.edu,  <https://orcid.org/0000-0001-7312-9356> (YL); gautam-pant@uiowa.edu,  <https://orcid.org/0000-0001-7414-2325> (GP); olivia.sheng@business.utah.edu (ORLS)

Received: October 31, 2018

Revised: October 11, 2019; February 4, 2020;
May 4, 2020

Accepted: May 29, 2020

Published Online in Articles in Advance:
November 9, 2020

<https://doi.org/10.1287/isre.2020.0954>

Copyright: © 2020 INFORMS

Abstract. Human capital is a key component of the knowledge economy that firms compete for in the labor market. Compared with the product market competition, the identification and prediction of labor market competitors have garnered little attention in the literature. In this study, we perform an interfirm labor market competitor analysis with a unique longitudinal employer-employee matched data set derived from online profiles of 89,943 employees, tracking their careers in 3,467 public firms from the years 2000 to 2014. Using employee migrations across firms, we derive and analyze a human capital flow network. We leverage this network to extract global cues about interfirm human capital overlap through structural equivalence and community classification. The online employee profiles also provide rich data on the explicit knowledge base of firms. In particular, they allow us to represent firms in the space of the skills possessed by their employees and measure the interfirm human capital overlap in terms of similarity in their employees' skills. We validate our proposed human capital overlap metrics in a predictive analytics framework using future employee migrations as an indicator of labor market competition. The results show that our proposed metrics have superior predictive power over conventional firm-level economic and human resource measures. We also demonstrate how our proposed metrics and the prediction framework can be incorporated into a comprehensive competitor analysis that includes both product and labor overlap between firms.

History: Ahmed Abbasi, Senior Editor; Huimin Zhao, Associate Editor.

Supplemental Material: The online appendices are available at <https://doi.org/10.1287/isre.2020.0954>.

Keywords: competitor analysis • human capital • text mining • network analysis • machine learning

1. Introduction and Related Work

The current economy is characterized by the growing intensity in interfirm competition. In particular, firms not only compete for consumers in the product market but also compete for human capital in the labor market (Markman et al. 2009). The labor market competition between firms is important because human capital is key to firm success (Grant 1996a, Beechler and Woodward 2009). In theory, following the resource-based view (RBV) of the firm (Barney 1991), human capital has been identified as an important resource for a firm to establish and sustain competitive advantage (Wright et al. 1994). In practice, since the late 1990s, with the publication of McKinsey's "war for talent" research (Chambers et al. 1998, Axelrod et al. 2001), business executives have increasingly emphasized the need for their firms to effectively attract, motivate, develop, and retain talent (Wright and McMahan 2011). This is particularly true in the current knowledge economy, where far

more of a company's net worth is tied up in employees' knowledge than in tangible assets (Cliffe 1998). Also, the rapid changes in technology make it more difficult to adequately train and develop employees to meet firms' demands for talent, and they increasingly rely on the acquisition of human assets from other firms to satisfy their human capital needs (Cappelli 2008). Moreover, firms often hire to learn from other firms (Song et al. 2003). The literature has documented many high-profile cases of such competition between Walmart and Amazon (Gardner 2005), Tesla and Apple (Higgins and Hull 2015), and Wall Street and Silicon Valley (Wigglesworth 2015). The latter examples show that the labor market competition does not exist just within a product market or even within an industry but can span a diverse set of firms across industries. In other words, what makes the labor market competition even more important, interesting, and complex is the potential for varied implications for firms when the overlap between firms in

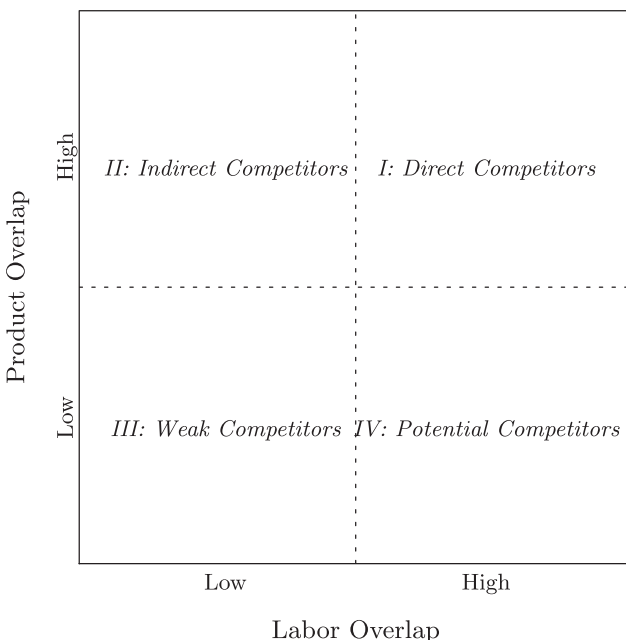
product and labor is taken into account simultaneously (Markman et al. 2009).

Research in business strategy has long recognized the fact that competitors can be identified not only by similarities among their products (market) but also by similarities among their resources. Hence, they have proposed theoretical frameworks of competitor analysis that integrate firm competition based on both the product and resource similarities (e.g., labor similarity), as shown in Figure 1 (Chen 1996, Peteraf and Bergen 2003, Markman et al. 2009). Drawing from this literature and applying it to a specific resource (i.e., labor), we can place each firm pair into one of the four quadrants of Figure 1. Each quadrant signifies different levels of similarities in the product and labor dimensions. For example, following the terminology of Bergen and Peteraf (2002), Amazon and Walmart may be categorized as *direct competitors* in the first quadrant of Figure 1, which provides a rationale for the strong reactions (filing a lawsuit) of Walmart when losing its information technology (IT) and logistics professionals to Amazon.¹ More generally, by adding the dimension of *labor overlap*, the competition framework shown in Figure 1 allows a focal firm to differentiate between product market competitors applying similar or dissimilar human capital endowments (quadrant I versus quadrant II). In other words, the *indirect competitors* in the second quadrant are likely to be firms that use different technologies for the production of similar products, which leads to the differences in their human capital needs. Such differentiation is important because introducing new resources to a particular production process is one of

the most effective ways for an entrant to disrupt an established product market (e.g., electric cars versus gas cars, digital cameras versus film cameras, synthetic diamonds versus natural diamonds; Markman et al. 2009). Similarly, the labor overlap dimension allows a firm to identify non-product market competitors that possess similar labor skills (quadrant IV) and hence pose a much stronger potential threat to the focal firm than firms in quadrant III (Bergen and Peteraf 2002). Certain firm pairs from Wall Street and Silicon Valley may be put into the fourth quadrant as *potential competitors* because they produce different products and yet may possess a similar set of human capital. In fact, the firm's human capital endowment is an important factor for its product development and diversification across industries (Farjoun 1994), and one may argue that the *potential competition* faced by the finance sector from big technology companies may have already been realized.² Again, we note that it is the incorporation of the labor market dimension of competition that differentiates the potential competitors in quadrant IV from the weak competitors in quadrant III (and direct competitors from indirect competitors) that may look the same if only the product similarity is considered yet posit very different strategic implications for the focal firm. Therefore, this two-dimensional (2D) competitor analysis framework poses great promise in providing valuable strategic insights for interfirm competition analysis (Chen 1996, Peteraf and Bergen 2003, Markman et al. 2009).

To empirically perform the 2D competitor analysis for a large number of firms and thus realize its potential practical utility, however, require the identification of interfirm similarities along both the product and labor dimensions. The product market similarity can be measured by, for example, the firm-pair similarity in their industry classifications or business descriptions (Shi et al. 2016). Pant and Sheng (2015) further use firms' websites and their hyperlinks to predict product market competitors. Also, commercial firm profiling companies such as Hoover's and Mergent manually identify product market competitors. However, previous works have shown them to be quite incomplete (Ma et al. 2011), probably because of the scale and dynamic nature of the problem (Pant and Sheng 2015). In contrast, despite its strategic value, empirically identifying firm labor market competitors has drawn much less attention from both academia and industry. We speculate that this lack of empirical studies stems from the unavailability of relevant data sources (and the underuse of many data science methods) that allow us to operationalize firms' behaviors in the labor market, such as employment (Horton and Tambe 2015) and human capital endowment (Wright and McMahan 2011). Moreover, as compared

Figure 1. Two Dimensions of Interfirm Competition



with product markets, the identification of competitors in the labor market can be more challenging because of cross-industry migrations of human capital. Hence, a predictive framework that can mitigate this challenge is beneficial for stakeholders both within a firm [e.g., human resources (HR) managers] and outside the firm (e.g., investors, analysts). In this paper, we take on a first of its kind endeavor to propose human capital overlap metrics derived from employees' skills and career mobility patterns across firms, which we use to predict future labor market competition.

As one of the most widely accepted theoretical perspectives in the strategic management literature (Barney 1991, Newbert 2007), the RBV of the firm provides a theoretical foundation for predicting labor market competitors with human capital overlap metrics between firms. Specifically, in RBV, firms are seen as resource bundles, and as long as these bundles are valuable and difficult to imitate or substitute, they can provide a firm with a competitive advantage (Hoopes et al. 2003). In other words, RBV highlights that resource bundles at a firm determine its competitive positioning. Human capital, being one of the most important resources of a firm (Barney 1991, Bartlett and Ghoshal 2002, Davenport et al. 2010), is hence critical to the firm's competitive positioning. In particular, Wright et al. (1994) provide a theoretical assessment of how human capital meets the criteria for sustained competitive advantage under RBV. Importantly, Wright et al. (1994) argue that what makes a firm's human capital difficult to imitate or substitute is the fact that it is a combination of its individual employees' human capital. Hence, even though individual employees of a focal firm can be imitated or substituted (e.g., with technology) by other firms, it is much more difficult to do so when the entire human capital pool of the focal firm is considered (e.g., the particular distribution of skills at the firm). Moreover, in addition to the explicit knowledge and skills, one important aspect of human capital is the implicit knowledge (Grant 1996a) held by employees. By definition, this tacit knowledge is difficult and slow to codify and hence imitate across firms (Grant 1996b). In other words, the RBV literature has illustrated that the composition of a firm's human capital in terms of the mix of knowledge and skills (both explicit and tacit) can characterize the nature of its human capital bundle. The exact composition of the human capital bundle can be expected to vary between firms, giving each firm a unique competitive positioning. Using this theoretical argument, if we can somehow quantify and measure the overlap in the composition of human capital at two firms, we can measure the similarity in their competitive positioning. Specifically, the composition can be viewed from two different perspectives: what do

the employees explicitly know (Section 3.1), and where do the employees come from (Section 3.2)? The second perspective can capture the tacit knowledge (e.g., work culture, managerial know-how) that the employees may bring to a firm. Therefore, after we can characterize the human capital bundle from these two perspectives, we can identify their competitive positioning through human capital. Then we can measure the overlap between firms in their competitive positioning to predict future labor market competition. In other words, our metrics provide a clear and rich operationalization of human capital configuration and its overlap between firms. Also, through RBV, we have a theoretically guided expectation that our proposed metrics will provide value in predicting labor market competition.

Empirically, a fundamental challenge in identifying labor market competitors is the measurement of human capital at the level of an individual firm, as described earlier (Wright and McMahan 2011). Previous studies have used surveys (Takeuchi et al. 2007, Ployhart et al. 2009) to measure properties of human capital at a firm, but they lack the necessary scale and granularity. As a result, the extant studies fail to identify knowledge and skills of individual employees beyond general assessments by their managers (Takeuchi et al. 2007) or through proxies such as level of education and work experiences (Hitt et al. 2001). More important for this study, existing human capital measures are not suitable for identifying labor market competitors because they do not provide relevant signals for that purpose. For example, it would be hard to argue that two firms are labor market competitors simply because they have similar numbers of employees with graduate educations or their employees have similar cognitive abilities. In other words, to analyze the competition between firms for human capital, a more detailed and richer measurement of human capital is warranted.

We address the human capital measurement problem with a unique longitudinal employer-employee matched data set from the information of more than 89,000 LinkedIn users' public profile pages. Typically, employees of firms do not maintain dedicated sites, but they have a substantial web presence through LinkedIn profiles, blogs, news stories, and so on. In particular, the publicly available LinkedIn profiles of employees (e.g., available through search engines) serve as a rich source of data to study the interactions between firms and employees over time (Tambe and Hitt 2011). These public profiles may provide information not only on the connections between firms and human capital but also on the nature of this capital in terms of experiences, education, and skills. As a result, we can track an employee's job history across different firms over time and hence locate a set of

employees working in a firm in a particular year. Moreover, because LinkedIn users typically report a set of *skill terms* to indicate the human capital they possess, we can aggregate the individual skill terms at the firm level. In this manner, we can construct a *skill summary* for each firm. For example, Table 1 shows the top-10 skills for some of the firms in 2011, where each skill term is weighted by the number of employees at the given firm that reported it. The firm-level skill summaries, as shown in Table 1, can serve as a granular description of the explicit knowledge base of a firm. In other words, compared with the popular firm-level human capital measures in the literature, Table 1 has the advantage of measuring human capital at the individual level and then representing each firm as a bundle of its employees' skills or knowledge. Hence, it contributes to the human capital literature by providing an important human capital measurement that closely matches the notion of RBV. In particular, we represent a firm as a bundle of weighted skills where the exact bundle (or the distribution over skills) is expected to be somewhat unique to the firm. Hence, our skill-based measurement of firms' human capital could be applied more generally to empirically test other human capital and firm strategy theories (Wright et al. 1994, Newbert 2007). In this study where we focus on interfirm competition for human capital, the firm-level skill summaries (as shown in Table 1) allow us to develop labor overlap metrics between firms much more effectively than what traditional measures such as education attainment provide. We thus suggest the use of interfirm labor overlap metrics based on employee skills in Section 3.1.

A second and more critical challenge in labor market competitor identification is the relatively little prior information or the ground truth about the firms' competitor relationships in the labor market. This is different from the product market competition, where firms' output-side information is available from a variety of sources, and firm pair similarity in products can be calculated and validated accordingly (Pant and Sheng 2015, Shi et al. 2016). There is one source (although not

readily available until recently) that can reveal labor market competition: employee career data. The matched employer-employee observations can provide information on the movement of employees between firms over time. Such individual movements, when aggregated over firms and time, can provide observations of labor market competition. Hence, we use the movement of employees (Gardner 2002, 2005) between two firms, which we call human capital flow (HCF), as the labor market competition measurement between them. We do recognize that the observed individual employee movement from one firm to another can be a result of many overlapping determinants (see Hom et al. 2017 for a recent review of the employee turnover literature). However, at the firm level, the fact that one firm hires employees from the other firm (regardless of the exact reasons for such hiring) indicates that a piece of the human capital (which includes explicit and tacit knowledge) available in the source firm is needed by the target firm of the observed HCF. Hence, using HCF as an indicator of labor market competition, we validate our proposed interfirm human capital overlap metrics using a predictive analytics framework (Shmueli and Koppius 2011) and evaluate how the metrics can provide value in predicting future HCF between firms.

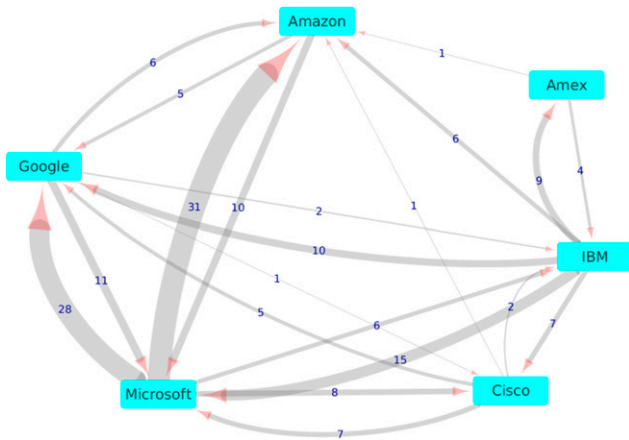
We connect the firms that have direct HCF between them to obtain an interfirm network structure that we call an *HCF network*. We conceptualize the HCF network, with the direction of the labor movement from a source to a target firm, as a *supply chain network of human capital* (Cappelli 2008). Figure 2 shows an HCF network based on a small subsample of our data to illustrate the notion of a supply chain network. In such a network, a given firm (e.g., Amazon) can be seen as receiving human capital from upstream firms (e.g., Microsoft, IBM, Cisco, American Express) and providing human capital to downstream firms (e.g., Microsoft, Google).³ When employees move from a source firm to a target firm, they bring not only explicit knowledge (e.g., Six Sigma) but also tacit knowledge (e.g., work culture, procedures) to the target firm

Table 1. Top-10 Skills of Example Firms in 2011

Amazon	Walmart	Ford Motor Company	IBM
e-commerce (75)	Customer service (390)	Automotive (384)	Cloud computing (216)
Management (71)	Retail (257)	Manufacturing (297)	IT strategy (195)
Customer service (68)	Inventory management (228)	Six Sigma (247)	Integration (192)
Java (65)	Microsoft Office (221)	Vehicles (227)	Solution architecture (183)
Software development (64)	Merchandising (200)	FMEA (219)	Project management (179)
Distributed systems (62)	Microsoft Word (180)	Continuous improvement (218)	Business process (175)
Leadership (61)	Microsoft Excel (168)	Lean manufacturing (195)	Program management (175)
Agile methodologies (57)	Cashiering (158)	Automotive engineering (181)	Business analysis (171)
Microsoft Office (55)	Time management (158)	PPAP (173)	Management (157)
Process improvement (54)	Inventory control (157)	APQP (151)	Enterprise architecture (153)

Notes. FMEA, failure mode and effects analysis; PPAP, production part approval process; APQP, advanced product quality planning.

Figure 2. (Color online) Illustration of an HCF Network Based on a Small Subsample of Data



Notes. The arrows show the direction of labor movement between firms. The width and the label of an arrow are indicative of the number of employees observed to move between the firms.

(Grant 1996b). Although skill summaries can capture the explicit knowledge, the network-based representation of firms in the labor market provides an opportunity to also capture the tacit knowledge flows. If firms are seen as human resource bundles (RBV), a firm's human capital can be characterized by its HCF network relationships because these relationships signify (explicit and tacit) knowledge flows. Hence, a similarity between two firms in terms of upstream relationships reflects a demand for and flow of similar human capital to the firms. Likewise, a similarity between two firms in terms of downstream relationships reflects similar human capital at the firms (as perceived by the labor market). Because HCF network relationships indicate unique human resource configurations of firms, based on RBV, we can expect them to also provide cues to similarities in firms' competitive positioning.

The network representation of firms through the employee migrations between them in the literature is rare. In particular, Guerrero and Axtell (2013) and Schmutte (2014) have both shown that conventional industrial classification is a poor signal of clusters derived from employee migration networks. At the firm level, structural properties in such networks have been shown to be related to firm performance (Wu et al. 2018). However, previous research has not used such networks or their properties for predicting labor market competition. We observe significant predictive power of network-based human capital overlap metrics in addition to the skill-based human capital overlap metrics. The additional information contained in network-based metrics may indicate the need and form a basis for the development of new theories regarding the interfirm competition for human

capital (Shmueli and Koppius 2011). For example, new theoretical development may involve investigating conditions that fuel the competition for different types of human capital (e.g., explicit versus tacit knowledge) as well as documenting the differences among competitors (e.g., from similar or dissimilar product markets) for different types of human capital (Chen and Miller 2012).

In summary, the previous literature (Chen 1996, Peteraf and Bergen 2003, Markman et al. 2009) has identified a theoretical framework for analyzing pairwise competition between firms through their overlap on the product market and resource dimensions (Figure 1). Also, RBV suggests that the characterization of firms' human resource bundles can allow us to identify similarities in their competitive positioning. However, these streams of work, although being theoretically clear, do not provide guidance on the empirical measurement of interfirm overlap along the human capital dimension. In response to this research gap, this study provides the following main contributions:

- We propose novel human capital overlap metrics based on firms' skill endowment and their embedded HCF network structure (Section 3). Although the human capital overlap based on skill endowment allows for capturing the interfirm similarities in the explicit knowledge base, the overlap in terms of the HCF network structure allows for capturing similarities on a broader scope of both the tacit and explicit knowledge base. Grounded in the resource-based view of the firm, our proposed human capital metrics could be applied more generally to empirically test other human capital and firm strategy theories and derive business intelligence (Wright et al. 1994, Newbert 2007).

- This is the first study on predicting labor market competition, and we use a wide variety of economic, human resources, product overlap, and human capital overlap metrics as predictors (Section 5). We tease out the predictive utility of these different types of metrics with a focus on additional utility provided by the proposed human capital overlap metrics. We experiment with a number of state-of-the-art machine learning methods ranging from logistic regression with regularization to random forest and deep learning.

- By tackling the empirical measurement of human capital overlap and validating it through a predictive analytics framework, we can operationalize the 2D theoretical framework (Figure 1) into an empirical lens through which strategic insights can be derived about interfirm competition by managers and analysts (Section 6). We present this operationalization at various levels of granularity (e.g., all firms, industries, individual firms). Given the key role of human capital for firms to succeed in the knowledge economy,

the operationalization of the 2D competitor analysis along with the prediction of firm labor market competitors benefits both the managers within the focal firms and those in external firms.

- We also provide a previously unexplored network analysis based on employee migrations that shows the small-world nature of the HCF network with weak industrial homophily that quickly diminishes with link distance (Section 3.2). This is clear evidence on the boundarylessness of employee careers as they migrate across industries. It also highlights the need for network-based metrics because such metrics can provide a more global view of human capital overlap.

In the next section, we describe in detail our unique employer-employee matched data set derived from public profiles of employees.

2. Data

We seed our data by focusing on employees of Standard & Poor's 100 Index (S&P 100) companies as of May 2015. One reason for choosing the S&P 100 companies as our seeds is that they include companies across multiple industry groups, thus leading to a diversity of firms in our data set. If we were to randomly sample all the LinkedIn users (although this is not practically possible because we do not have access to all of LinkedIn data), the resulting sample would be greatly biased in terms of industries that are overrepresented on LinkedIn (e.g., technology firms).⁴ In addition, because S&P 100 companies represent the most valued companies (by the market), their employees (on average) would tend to represent the more highly valued human capital than the employees from a random group of firms. Identifying labor market competitors for such human capital would be of greater utility.

We obtain the publicly available LinkedIn profiles of employees of S&P 100 firms by searching through the Yahoo BOSS application programming interface, which provides programmatic access to Yahoo web search data. In particular, we search for pages from linkedin.com that pertain to a given S&P 100 firm. We use a combination of keywords that have a high probability of identifying public profile pages of employees of the firm with information on their job experiences, education, and skill terms. We repeat the process for all S&P 100 firms and obtain up to 1,000 such profiles for each firm.⁵ As a result, we obtain the LinkedIn profiles of 89,943 individuals who are likely to be working in S&P 100 companies as of May 2015.

A typical publicly available LinkedIn profile in our data set is similar to a brief curriculum vitae. It contains an employee's job experiences, education information, and skill terms. In particular, the job experience of an individual includes the firm name

and the start and end dates of this job experience. Based on all individuals' job experiences, we identify 75,350 different companies/organizations for which the employees had worked at some point in time. We obtain other firm-level data such as revenue, firm size (number of employees), and business description from the Compustat North America database, and only firms whose information can be found in the Compustat database are included.⁶ As a result, by seeding employees from S&P 100 firms and using their past work experiences, a total of 3,467 publicly held firms are included in our analysis. There are around 5,000 publicly traded companies in the United States (Doidge et al. 2017); hence, our coverage of firms is substantial.

From an individual's job experiences, we can observe where the individual was working (among the 3,467 firms included in our study) in a particular year. As a result, we compile a longitudinal employer-employee matched data set that contains a large number of employees and firms and a relatively long period from 2000 to 2014. Such detailed information from the matched employer-employee records is essential for our labor market competition study but unavailable from most public data sources. In addition to seeding our data from a diverse set of firms, we perform robustness checks to verify the representativeness of our data. Because ours is a firm-level study, we compare basic firm-level statistics from our data with all the firms in the Compustat database. Table 2 shows the percentage of firms in each of the one-digit Standard Industrial Classification (SIC)-level industries. We observe that except for the finance sector (SIC Code = 6), the distribution of firms across different industries is similar between firms in our data and firms from Compustat. The average absolute value of the differences (in percentages) between firms in our data and firms from Compustat is 4.17% (2.85% when the finance sector is not considered). Hence, our data include firms across all major industry groups, and their distribution across industries is similar to all firms included in the Compustat database. We also test how our sample matches with all employees at different firms over time. We compute the Kendall rank correlation between the sizes of firms (in the number of employees) based on our sample with the known sizes of the firms (obtained from Compustat). Figure 3 shows the Kendall rank correlations over time, which are consistently significant (at the 0.1% level). The rank correlation values confirm that our data are a reasonable representation of firms in terms of their relative sizes. We present additional analysis in terms of the representativeness based on employee skills and business summaries of firms in our data, as described in Online Appendix B.

Table 2. Percentage of Firms in Each Industry at One-Digit SIC Code Level

Industry	SIC Code	Firms in our data	Firms from Compustat
Agriculture, forestry, and fishing	0	0.21%	0.30%
Mining and construction	1	8.00%	11.58%
Manufacturing	2	15.83%	11.91%
Manufacturing	3	21.61%	16.13%
Transportation and public utilities	4	10.88%	7.68%
Wholesale and retail trade	5	10.32%	6.27%
Finance, insurance, and real estate	6	12.78%	28.79%
Services	7	16.31%	12.97%
Services	8	3.65%	2.79%
Public administration and nonclassifiable	9	0.42%	1.58%
Total percentage		100%	100%
Total number of firms		3,467	23,271

Workers in our data are highly educated: about 90% of workers in our data report a college or higher level of education. Therefore, the competitor analysis we perform in this paper should be seen as a competition between firms for the highly skilled/educated workers, which is important by itself because high-skilled employees are crucial to firm success (Grant 1996b, Beechler and Woodward 2009).

Also, as an online professional social network, one of the main motivations for an individual to have a LinkedIn profile is career enhancement or job search. Compared with a random sample of workers, those who have LinkedIn profiles are more “active” in the labor market (e.g., it is estimated that more than 85% of employed job seekers look for jobs online; Kuhn (2014)). Also, our search query is likely to find LinkedIn users with complete information on job experiences, education, and skill terms. In other words, workers in our data are more likely to be the target of labor market competition between firms. As a result, compared with a random sample of workers, our data are more likely to capture the employee mobility between firms, which is an important indicator of

interfirm labor market competition. We would also like to highlight the lack of alternate data sources that track employee-firm interactions in the labor market as well as employee skills, which is likely the main reason for the lack of empirical studies on interfirm labor market competition.

3. Human Capital Overlap Metrics

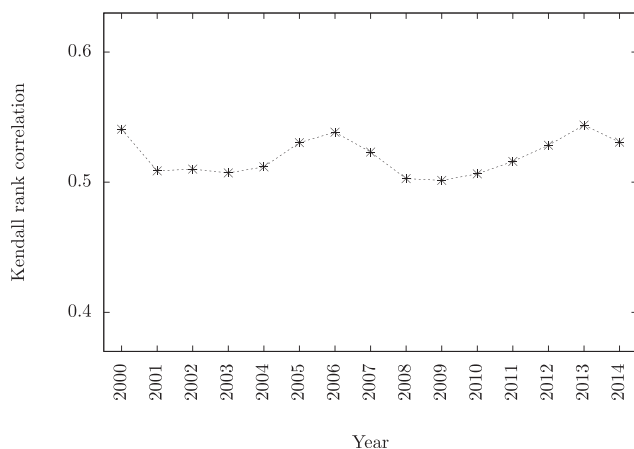
An overlap in the human capital of a pair of firms can be expected to be a leading predictor of their labor market competition (Peteraf and Bergen 2003). Based on this conjecture, we propose metrics that represent firm pair similarity in terms of their human capital. One way to measure the human capital overlap between a pair of firms is to view it through the lens of skills possessed by their employees. If two firms have employees with similar skills, they can be seen as having a human capital overlap in terms of explicit knowledge. We call this type of skill-based human capital similarity *labor overlap*. We suggest two related but different ways of measuring labor overlap.

Employees, through their work at firms, also gain a considerable amount of intangible or tacit knowledge (e.g., work culture, industry best practices, procedures, etc.; Ambrosini and Bowman 2001) that is not captured through self-reported skill terms (Hatch and Dyer 2004, Wright et al. 2014). However, the network induced by employee migration between firms (i.e., the HCF network) could provide cues to the human capital overlap that is broader than the explicit skills alone. For example, if two firms hire from a similar set of upstream firms, they can be expected to import similar tacit knowledge from the source firms. We suggest several *HCF network overlap* measures that attempt to capture this broader notion of human capital overlap.

3.1. Labor Overlap

Of the 89,943 individual employees included in our data, 86,030 of them have reported skill terms. All skill

Figure 3. Correlation Between Firm Sizes from Our Sample and True Firm Sizes



terms that are reported by more than one employee are included in our analysis.⁷ Hence, a total of 15,998 distinct skill terms are used in all the subsequent analyses. On average, an employee reported 14.6 skill terms. The skill terms reported by individual employees form the basis for our construction of inter-firm skill-based similarity metrics.

3.1.1. Skill Term Similarity. As described earlier, we construct a skill summary for each firm by aggregating the skill terms of its employees in a particular year (see Table 1). This firm-level skill summary can be seen as an operationalization of the theoretical notion of *human capital pool* that signifies the aggregated skill base of firms (Wright et al. 2014) and serves as a description of the explicit knowledge and expertise that are embedded in the firm's human capital for a given year. Specifically, we represent each firm k as a *skill vector* \mathbf{s}_k in \mathcal{R}^N space, where $N = 15,998$ is the set of all skill terms across employees in our sample. Similar to the idea of term frequency-inverse document frequency weighting in information retrieval (Manning et al. 2010), each element of \mathbf{s}_k is computed as the product of skill frequency SF and inverse firm frequency IFF

$$SF \cdot IFF_{s,k} = SF_{s,k} \times IFF_s. \quad (1)$$

The skill frequency $SF_{s,k}$ is defined as the number of employees at firm k that list the skill term s in their LinkedIn profile. The inverse firm frequency IFF_s is defined as $\log \frac{F}{FF_s}$, where F is the total number of firms, and FF_s is the number of firms whose skill summary contains the skill term s (i.e., they have one or more employees with skill s). The inverse firm frequency helps to weigh down skills that are common and hence, appear in profiles of employees across many firms. We would like to note that a skill term with a higher value of IFF does not mean that the skill is more important. However, IFF weighting could indicate specific versus general skill terms and hence allow us to differentiate firms with respect to their skill summaries better. For example, skills such as "Microsoft Office" are perhaps prevalent among employees in most organizations, whereas skills such as "distributed systems" can be specific to certain firms or sectors. While considering the similarity between two firms, being similar in general skills is likely to be less informative about their human capital overlap than being similar in specific skills.

Given that \mathbf{s}_k represents the skill term distribution at firm k , we can measure the similarity in the human

capital at two firms by measuring the cosine of the angle (i.e., cosine similarity) between the skill vectors corresponding to firms a and b as follows:

$$\text{sim}(a, b) = \frac{\mathbf{s}_a \cdot \mathbf{s}_b}{\|\mathbf{s}_a\| \cdot \|\mathbf{s}_b\|}. \quad (2)$$

We call this the *skill term similarity* (*SkillTermSim*) between firms. The cosine similarity-based metric has the advantage that it measures the similarity in the relative distribution of skills between the two firms and hence normalizes the effect of the size of the firms (in the number of employees).⁸

3.1.2. Skill Topic Similarity. The skill term similarity, as just presented, treats each skill term independently. However, different skill terms can reflect the same overall types of skills or human capital. For example, the skill terms "java," "c++," and "c#" can represent the more general skill of programming language, and the skill terms "project management," "program management," and "leadership" can all represent general management skill. Motivated by such observations, we apply the unsupervised latent Dirichlet allocation (LDA; Blei et al. 2003) to discover the underlying *skill topics* in the employees' reported skill terms. We choose the LDA because it has been applied successfully to classify various types of documents, including pictures, scientific articles, social network data, and survey data (Blei 2012). Given the unsupervised nature of such tasks, the LDA has been widely used for identifying latent groups that can be subjected to further analysis. In particular, Shi et al. (2016) apply the LDA to discover the latent topics in firms' business descriptions and construct a firm pair proximity measure based on firms' similarities in the space of the discovered business topics.

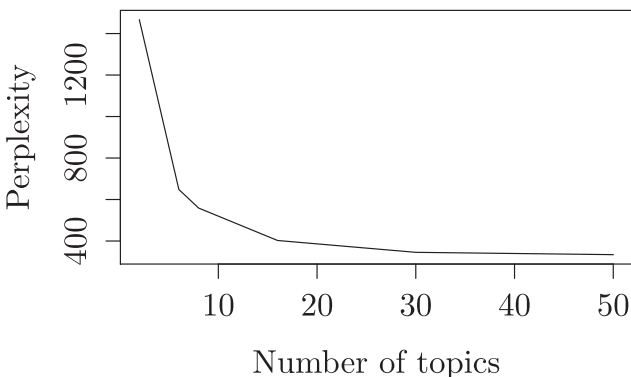
More specifically, the LDA algorithm assumes the following generating process for a corpus of text documents. Each document is represented as a mixture over a small number of latent topics, where each latent topic is characterized by a distribution over all the words, and each word's appearance in a document is attributable to one of the document's latent topics. In our context, we treat the skill terms reported by an individual employee as a single *document*. Then all the 86,030 employees' skill terms can be represented by a document-term matrix with 86,030 rows and 15,998 columns, where each element in this matrix is a binary indicator of whether an employee (row) reported a particular skill term (column). The LDA output represents each employee with a probability distribution over a set of the latent skill topics,

where each skill topic is represented by a probability distribution over all 15,998 skill terms. As a result, given an employee's skill topic distribution, we can identify the skill topic to which he or she most likely belongs. In this manner, we can classify all employees into different skill topics.

The input parameter for the LDA algorithm is the number of different latent topics. To decide this number, we follow Blei et al. (2003), experiment with different values, and calculate the in-sample *perplexity* for each number of topics. The perplexity value decreases with the number of topics chosen, and a smaller perplexity value indicates a better fit of the model (Blei et al. 2003). By contrast, a smaller number of topics is preferable to avoid overfitting. Figure 4 plots the perplexity value of the LDA model for a different number of topics. Based on this figure, we set the number of skill topics as six because it provides semantically meaningful topics and proceed with this value for all subsequent analyses.⁹

The LDA algorithm represents each of the six skill topics with a distribution over the 15,998 skill terms. Figure 5 shows the top-10 skill terms with the highest probabilities for each of the six skill topics. The horizontal axis in each subplot represents the probability of a skill term appearing in a skill topic. For example, we observe that the lower center subplot in Figure 5 shows a skill topic that is characterized by high probabilities for skill terms related to IT (the most likely skill term for this skill topic is “java,” with a probability of 0.015). We also note that the other five topics have a different set of top skill terms compared with the IT skill topic, and each skill topic can represent a recognizably distinct skill category. Based on the top skill terms for each skill topic category, we refer to category 1 as “Analyst-Admin,” category 2 as “Biotech-Health,” category 3 as “Production-Operation,” category 4 as “Sales-HR,” category 5 as “IT,” and category 6 as “Management.”

Figure 4. Perplexity Plot for the Number of Topics Based on All (86,030) Employees



For each of the 86,030 employees, the LDA algorithm assigns a probability distribution over the six skill topics. In other words, each employee is represented by six probability values corresponding to the six skill topics. Each firm d in a year can then be represented by a vector θ_d of size six, where each element of the vector is the sum of its employees' probabilities for that skill topic. As a result, for two firms a and b , we calculate the interfirm similarity based on the skill topic distribution of their labor as

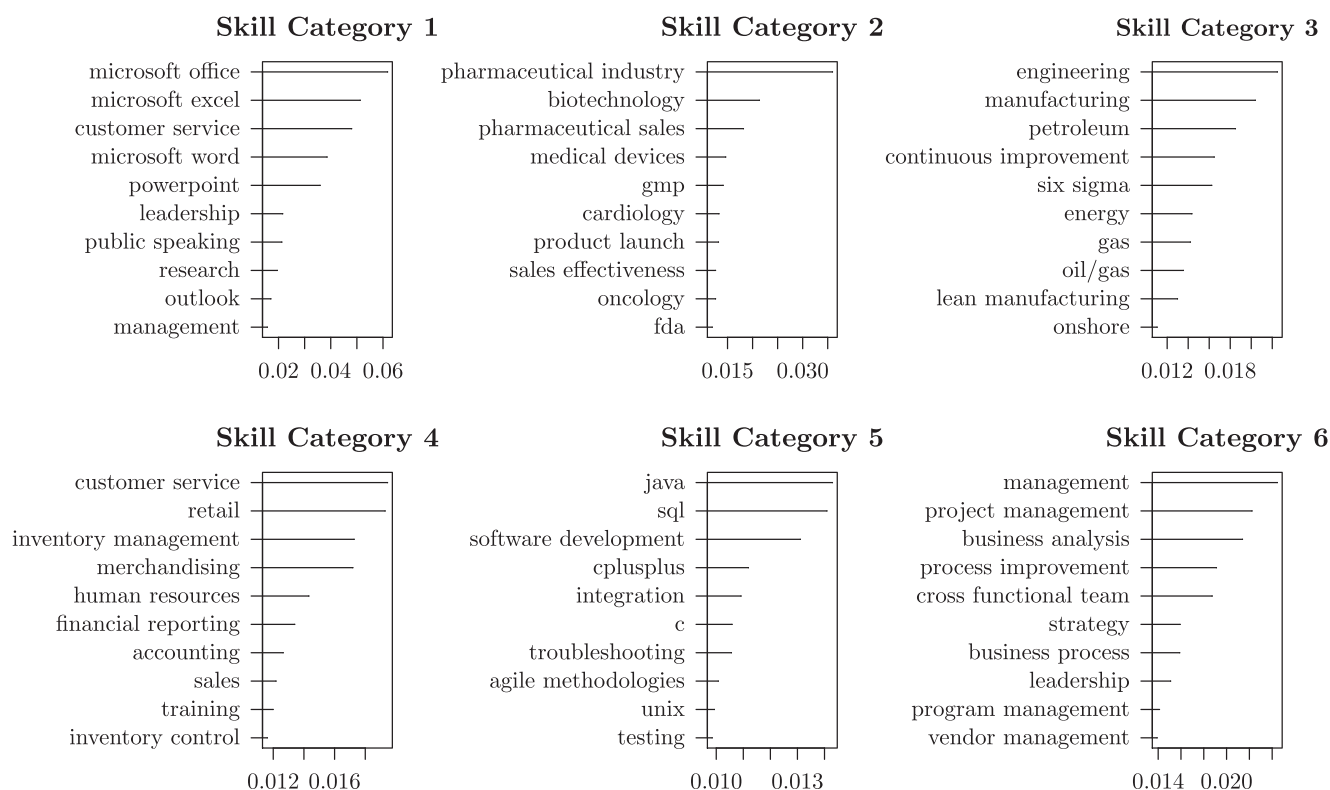
$$\text{sim}(a, b) = \frac{\theta_a \cdot \theta_b}{\|\theta_a\| \cdot \|\theta_b\|}. \quad (3)$$

We call this the *skill topic similarity* (*SkillTopicSim*) between firms.

The two labor overlap metrics are based on firms' endowments in employee skills and hence can be measures of input-side or resource-side proximity between a pair of firms. Shi et al. (2016) have proposed a firm proximity measure in terms of outputs (or products) by measuring the similarity in the business descriptions of firms. Our labor overlap metrics extend previous interfirm similarity research that has largely focused on the product-side overlaps. Although some effort has been made in the past to measure the similarity of human capital based on industry-level details (Farjoun 1994), we considerably extend these efforts as well by using individual-level skills. Such individual-level human capital data over a large number of firms would have been challenging and costly to obtain before the appearance of platforms such as LinkedIn. Moreover, by aggregating employee skill terms at the firm level, it is possible to track the firm's “human capital pool” over time (Wright et al. 1994, pp. 304).

3.2. HCF Network Overlap

The employee migrations between the firms included in our data allow us to connect the firms using a network structure. The nodes of the network are firms, and the directed edges between firms are weighted by the number of employees who have moved from the source firm to the target firm up to a given year (see Figure 2). In other words, two firms are connected in the HCF network by the direct employee mobility between them. For simplicity, we start by using an undirected and unweighted version of the HCF network to derive some summary network measures. Specifically, when all HCF over time is considered, the network contains 3,467 nodes and 20,171 edges. A large connected component exists where 3,465 firms are connected through paths. The network has a diameter of nine and an average shortest path of 3.14 edges. In other words, the HCF

Figure 5. Skill Topics Based on All (86,030) Employees

Note. The horizontal axis shows probabilities for the top-10 skill terms in each skill topic.

network shows *small-world connectivity*, where two firms, on average, are a little more than three links away, and the degree distribution is skewed. The small-world observation resonates with the “boundaryless career” idea proposed in the literature (Arthur 2014, pp. 627). Contrary to the organizational career, where an employee’s career evolves in the same organizational setting over time, a typical employee is not expected to spend his or her whole career in one organization in the current economy. Besides, the small diameter (and short paths) of the HCF network is also a reflection of the boundarylessness of employees’ careers in terms of industries. As an example, Figure 2 plots a subgraph of the HCF network where edge weights are the aggregated HCF until 2012. We notice both intraindustry and interindustry movements (between IBM and American Express) in this small example. Because employee migration between firms has been noted to be a source for institutional isomorphism (firms becoming similar over time; DiMaggio and Powell 1983, Pant and Sheng 2015), we can expect firms to seek counterbalancing hiring strategies to find human resources who bring new knowledge and hence help with institutional divergence (Beckert 2010). The dual forces of institutional isomorphism and divergence may explain firms’ willingness to hire from both within and across

industries. In other words, there are both demand-side (firms) and supply-side (employees) factors that can explain the small-world connectivity of the HCF network.

A network of firms can be constructed in many different ways. The information contained in a property of the network depends on the nature of how the firms are connected. For example, in a network where firms are connected by collaboration or alliance, knowledge transfer is more likely between connected than unconnected firm pairs (Almeida and Kogut 1999, Rosenkopf and Almeida 2003, Schilling and Phelps 2007). However, such a collaboration or alliance network is not expected to embed within it information on competition between firms whether it is demand-side competition (e.g., similar products) or resource-side competition (e.g., labor market). Because in this study our focus is on labor market competition, firms are connected by the known movement of employees between them. An observed job hop of an employee can be a result of complicated sequences of interactions between employees and employers. However, the flow of human capital from firm *a* to firm *b*, regardless of the reason, does imply that a piece of firm *a*’s organizational knowledge (tacit or explicit) is available to firm *b*. In particular, employee migrations capture the interfirm flow of tacit knowledge that is otherwise hard to encode. As a result,

the HCF network records which firms have acquired knowledge from which other firms over time. The structural properties of a firm in this network can be proxies for its *knowledge state* relative to other firms, both locally (direct neighbors) and globally (non-direct neighbors). More important, such properties can be hard to measure through firm-level variables alone or through explicit skill summaries. For this purpose, it is important to identify network-based variables that could provide complementary information on the interfirm human capital overlap.

3.2.1. Industrial Homophily in the HCF Network. We may expect that firms that are in similar industries are closer (based on link distance) on the HCF network, a phenomenon we call *industry locality*. If firms are in the same industry or similar industries, we may expect a greater likelihood that there is a direct link or a short indirect path between them on the HCF network. This is akin to what is called *value homophily* in social networks (McPherson et al. 2001) or *topical locality* on the web (Davison 2000, Pant and Srinivasan 2013).

Given that the SIC Codes follow a hierarchical structure with various granularities of industrial sectors indicated by their digits,¹⁰ we define SIC Code similarity (*SICSim*) between a pair of firms as shown in Table 3. We note that the SIC Code similarity metric that we suggest is more flexible and informative than the prevalent practice of using the first two or all the digits of the SIC Code to indicate a firm's industry (Weiner 2005).¹¹ After computing the SIC Code similarity of each pair of firms, we average the metric at various link distances. Figure 6 shows the average SIC Code similarity at various link distances on the HCF network (with undirected and unweighted edges). We find that firms that are one link away on the network have significantly higher SIC Code similarity than firms two links away, and they, in turn, have significantly higher SIC Code similarity than those three links away, and so on. We note that although the firms that are one link away do not show high SIC Code similarity (the metric in Table 3 varies between zero and four) to begin with, the existing industrial similarity drops quickly with link distance. In other

words, the HCF network shows industry locality and that it should be used in any subsequent model of employee migrations. However, it is important to note that the signal because of industry locality is weak, is localized, and drops quickly with link distance. Hence, it cannot be expected to account for a large part of the employee migration across firms. Despite this weakness, given the popularity of SIC Codes for identifying industries (Witten and Frank 2005), *SICSim* will serve as an important control variable for our study.

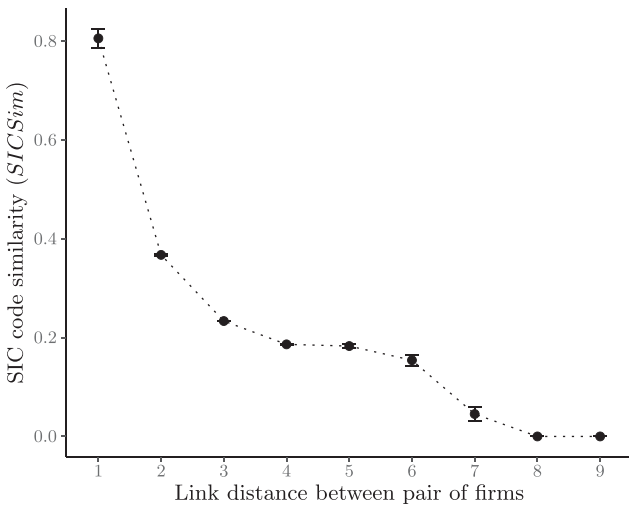
After observing some basic properties of the HCF network such as the small world and weak industrial localization, we can expect the network to provide complementary and global cues about labor market competition that are not captured by local firm-level variables. In particular, we suggest metrics that measure the level of HCF network overlap between a pair of firms. When two firms have an overlap in terms of their network neighborhood, we can expect them to have a similar human capital configuration and knowledge base (including the tacit kind), which can be expected to signal future employee migrations.¹² For all subsequent analyses, we will use weighted and directed HCF networks.

3.2.2. Upstream and Downstream Similarity. Given the network structure as described, for a firm pair in this network, we calculate their similarity in their upstream and downstream firms. The upstream firms of a focal firm are the firms from which employees have moved in the past to the focal firm. The downstream firms of a focal firm are the firms to which the focal firm's employees have moved in the past. Also, the links in the HCF network are weighted by the number of people who have moved between firms. Hence, we can represent each firm k as an upstream vector \mathbf{u}_k in \mathcal{R}^F space, where F is the set of all firm nodes in the network. Each element u_{ik} of \mathbf{u}_k is the number of employees who have moved in the past from firm i to firm k . In other words, \mathbf{u}_k represents the distribution of employees who have migrated to firm k over all firms. If each firm is seen as a unique knowledge stock (of explicit or tacit kinds), \mathbf{u}_k represented a

Table 3. Metric for Measuring SIC Code (Product) Similarity (*SICSim*) Between Firms

Type	Example	Similarity score
Digit 1 different	5411 and 6021	0
Digit 1 same, digit 2 different	5411 and 5072	1
Digits 1 and 2 same, digit 3 different	5411 and 5400	2
Digits 1–3 same, digit 4 different	5411 and 5412	3
Same four digits	5411 and 5411	4

Note. SIC Code 5411: retail-grocery stores; 5412: retail-convenience stores; 5400: retail-food stores; 5072: wholesale-hardware; 6021: national commercial banks.

Figure 6. SIC Code Similarity Between Firms at Various Link Distances in the HCF Network

distribution over all knowledge stocks represented by the F firms. In this sense, \mathbf{u}_k can be seen as a knowledge input into each firm that is acquired by the firm through incoming employees from other firms. Given that \mathbf{u}_k represents the knowledge input at firm k , we can measure the similarity in knowledge inputs to two firms a and b by measuring the cosine similarity between the upstream vectors \mathbf{u}_a and \mathbf{u}_b corresponding to the firms. We call this the *upstream similarity* between firms a and b .

Like upstream similarity, we can also compute the *downstream similarity* between a pair of firms. For this purpose, we represent each firm k as a downstream vector \mathbf{d}_k in \mathcal{R}^F space. Each element d_{ki} of \mathbf{d}_k is the number of employees who have moved in the past from firm k to firm i . Unlike the upstream vector, the downstream vector of a firm k represents the distribution of its outgoing employees over all firms. It represents the different degrees to which other firms value the human capital at a focal firm k . Again, if each firm is seen as a knowledge stock, \mathbf{d}_k represents a distribution of perceived utility of firm k 's knowledge stock for other firms. In other words, when a similar set of downstream firms hires from a pair of firms, it suggests that the pair of firms has similar knowledge stock (both explicit and tacit kinds) as perceived by other firms.

3.2.3. HCF Network Community. Given the weighted and directed HCF network in a year, we apply a network community detection algorithm to find firm communities or clusters in the network. An HCF network community is essentially a set of firms that have dense connections between them and sparser connections with other communities. Because connections indicate past employee migrations, we can

expect firms in the same HCF network community to have a greater likelihood of competing in the labor market. More specifically, we follow Schmutte (2014) and apply the modularity maximization (Blondel et al. 2008) algorithm to find network communities.¹³ One advantage of the modularity maximization algorithm is that it does not require the number of communities as an input and finds network community structures by searching over different numbers of communities. Also, the resulting modularity value of a network provides information in terms of the quality of the community detected because modularity is zero for a random network with no community structure and one for a completed segmented network. A modularity value greater than 0.3 generally indicates substantial community structure (Blondel et al. 2008), and our HCF network has modularity greater than 0.3 for all years. After identifying the network communities, we can detect whether a firm pair in a given year falls into the same community. Hence, we create an indicator variable (one if two firms are in the same community, zero otherwise) that records the network community overlap for each pair of firms. We note that our network features (upstream/downstream similarity and community overlap) are all at the level of a firm pair. This is because our focus is on the labor market competition between two firms. A firm pair has been noted to be the appropriate unit for competitor analysis (Chen 1996). Hence, we do not use the (firm) node-level network measures as, for example, used by Wu et al. (2018), who study individual firm productivity.

4. Control Metrics

A basic firm competitor analysis starts by assuming that each firm has a unique product market profile and resource (e.g., human capital) endowment. Hence, the competitive relationship between two firms can be illuminated by their overlap in these two dimensions (Chen 1996, Markman et al. 2009). This basic view suggests that a pairwise competitor relationship between firms in the product market is associated with their similarity on the product side (i.e., demand side; Pant and Sheng 2015). Following a similar argument, a pairwise competitor relationship between firms in the labor market is expected to be associated with their similarity in human capital endowment (Markman et al. 2009). In the preceding section, we proposed a set of metrics for the human capital overlap between firms based on their employees' skills as well as the network structure derived from the employee mobility pattern between firms. Although the former can capture employees' explicit knowledge, the latter can also capture their tacit knowledge. A direct indicator of current labor market competition between firms is the employee migrations between those firms

(Gardner 2005). We hence validate our proposed human capital overlap metrics by evaluating their ability to predict future labor market competition operationalized using employee migrations. Because our goal is to predict future labor market competition, we include a set of basic economic control features that are expected to cue such competition (e.g., firm size, growth, etc.) and have been previously used in the human resource management literature. Also, as a result of the potential connection between firms' product market and labor market competition (Markman et al. 2009), we include measures of firms' product-side similarity in terms of their business and industry as predictors. In summary, we calculate three sets of features for our predictive analysis: basic economic metrics, product overlap metrics, and human capital overlap metrics (Table 4). The additional predictive power provided by our proposed human capital overlap metrics over and above the basic economic and product overlap metrics is a clear validation of our proposed human capital overlap metrics.

4.1. Basic Economic Metrics

Variables that indicate the current economic state of firms and the previous year's number of employees moved between firm pairs (*HCF_{lag}*) can be expected

to provide information on the future labor market competition. Hence, we record for each firm its revenue, revenue growth rate, number of employees, and growth rate in the number of employees from Compustat. The revenue and size of a firm can be seen as an indicator of power and maturity in the labor market. Similarly, the corresponding growth could be viewed as an indicator of a firm's current dynamism. We note that firm labor market competition can be skewed by firm size (e.g., between large firms). It is hence important to include such metrics that control firm size and growth to evaluate additional predictive utility of our proposed metrics. We also calculate the net HCF (number of incoming employees minus number of outgoing employees) observed in our data set. In addition, for a source-target firm pair, we compute the inverse HCF (from the target to the source firm) in each year. Although we do not have sufficient guidance from the literature to determine the utility of these variables in predicting the labor market competition, they serve as a set of important control variables for our current investigation.

Employee migration between firms can be expected to be sensitive to the state of HR at those firms (Hom et al. 2017). We calculate several conventional HR metrics (Wright and McMahan 2011) that indicate the

Table 4. Summary of Predictors

Variable	Description	Mean	Standard deviation
Panel A: Basic economic metrics			
<i>HCF_{lag}</i>	HCF in the previous year	0.24	1.00
<i>InvHCF</i>	Number of employees who move from a target to a source firm	0.08	0.46
<i>NetHCF</i>	Number of incoming employees minus number of employees leaving a firm	−0.61/−0.73	9.49/15.95
<i>Rev</i>	Revenue of a firm (millions)	37.21/42.89	62.13/61.84
<i>RevGro</i>	Growth rate of the revenue of a firm	0.12/0.13	7.78/3.77
<i>Emp</i>	Number of employees of a firm (thousands)	97.1/ 99.07	196.0/189.1
<i>EmpGro</i>	Growth rate of number of employees of a firm	0.10 / 0.06	19.12/1.56
<i>AvgYearWorking</i>	Average number of years since bachelor's degree for the employees	6.09/7.35	5.06/4.32
<i>PctGraduate</i>	Percentage of employees with a master's or PhD degree	0.15/0.15	0.19/0.16
<i>AvgUniversityRank</i>	Average ranking of employees' bachelor's degree granting universities	194.0/184.0	30.9/23.8
Panel B: Product overlap metrics			
<i>SICSim</i>	SIC Code similarity between two firms	0.96	1.41
<i>BusdescTermSim</i>	Cosine similarity between two firms' business summaries	0.13	0.13
<i>BusdescTopicSim</i>	Cosine similarity between two firms' business summary topics	0.24	0.28
Panel C: Human capital overlap metrics			
<i>SkillTermSim</i>	Cosine similarity between two firms' skill summaries	0.19	0.23
<i>SkillTopicSim</i>	Cosine similarity between two firms' employee skill topic distributions	0.54	0.37
<i>UpstreamSim</i>	Cosine similarity between two firms' inlink HCF network neighbors	0.13	0.25
<i>DownstreamSim</i>	Cosine similarity between two firms' outlink HCF network neighbors	0.23	0.29
<i>HCFCommunitySim</i>	Whether two firms belong to the same HCF network cluster	0.53	0.50

Notes. Because our data include source-target firm instances, for individual firm variables (such as revenue), each observation contains a value of the source and target firms, respectively. For example, a source-target HCF firm pair includes the source firm revenue and the target firm revenue. The summary statistics of individual firm variables hence include both the source and target firm values with the format (source firm/target firm).

state of HR at a firm based on the education information available in public LinkedIn profiles of current employees of the firm. Specifically, we compute the average number of years current employees have been working after college (i.e., their undergraduate degree).¹⁴ This variable can be considered a proxy for the average age and experience of the employees in a firm. Also, we compute the percentage of employees with a postgraduate degree and the average rank of employees' undergraduate universities.¹⁵ Again, these variables are the most commonly used human capital measures employed in the literature (Nyberg et al. 2014) and can be expected to provide signals on the level and quality of education among a firm's employees. The economic control variables are summarized in panel A of Table 4.

4.2. Product Overlap Metrics

Motivated by the firm competitor analysis framework (Markman et al. 2009) as shown in Figure 1, we calculate metrics that represent two firms' overlap in terms of their products. Although product overlap is unlikely to be sufficient to capture all the HCF between firms (Peteraf and Bergen 2003), it can provide important control metrics. We start with the SIC Code of firms because SIC Codes are reflective of the product side of firms. Using the hierarchical structure of the SIC Codes, we have suggested the SIC Code similarity (*SICSim*) measure between a pair of firms in Section 3.2. The variable *SICSim* provides us with a metric of product-side overlap between firms based on their industrial sectors.

One limitation of the proposed SIC Code similarity between firm pairs is that the SIC Code of a firm may not reflect all the different and granular product spaces in which the firm operates. To overcome this shortcoming, Shi et al. (2016) apply the LDA algorithm to detect latent business topics from textual business descriptions of firms. We follow Shi et al. (2016) and apply the LDA algorithm to the firm descriptions that are available from the Compustat database. The procedure of choosing the number of topics is identical to that in Section 3.1, where we apply LDA to the employees' skill terms. We calculate firm pair (cosine) similarity in their business topics (*BusdescTopicSim*) and also the (cosine) similarity in the text terms in their business summaries (*BusdescTermSim*). Panel B of Table 4 summarizes our firm pair product market overlap measures. Panel C of Table 4 shows our proposed labor overlap and HCF network overlap metrics as described in Section 3. For prediction analysis, all predictor values are standardized to represent the number of standard deviations from their mean values.

5. Predictive Analysis

The employee mobility between firms is a key reflection of interfirm labor market competition (Gardner 2002, 2005). Hence, the target variable of interest for our predictive framework is *HCF* in a given year. We experiment with our proposed set of metrics for the prediction of future labor market competition, which is operationalized using *HCF* values. An important part of our experimentation methodology is to tease out the utility that skill- and network-based human capital overlap metrics provide above and beyond the control metrics.

5.1. Data Set Construction

The unit of analysis for our prediction task is the source-target-year firm pair. In other words, for the predictive analysis, we construct a panel-type data set where source-target firm pairs are tracked over time. The time span we consider is from 2000 to 2014, and we use variables from year $t - 1$ (including *HCF_{lag}*) to predict the labor market competition outcome variable (as defined) in year t . Hence, each observation in our data represents the HCF outcome in year t with all predictor variables in the year $t - 1$. To construct a data set suitable for the goal of this study, we include only firm pairs with previously observed HCF between them. In other words, among the 3,467 firms included in this study, we start tracking firm pairs since their first observed HCF. As a result, a total of 84,733 source-target-year firm pair instances (excluding observations with missing predictor values) are included in our analysis. This data construction procedure allows the incremental addition of new source-target firm pairs into our data set over time. Such first-time firm pairs can be considered as new labor market competitors. Hence, we also evaluate the performance of the predictive models specifically on the new labor market competitors as an additional validation of our proposed metrics.

5.2. Outcome Variable

The outcome variable of interest is an indicator of interfirm labor market competition. Based on previous literature (Gardner 2005), we transform the numeric *HCF* values into a binary interfirm labor market competition indicator (Y) depending on whether the *HCF* value between the source and target firms meets a threshold δ :

$$Y = \begin{cases} 0, & \text{if } HCF < \delta, \\ 1, & \text{if } HCF \geq \delta. \end{cases} \quad (4)$$

In other words, we focus on a classification problem (i.e., classifying firm pairs into competitors or non-competitors). This binary classification is consistent

with the academic literature and practice where firms are either seen as competitors or not (Peteraf and Bergen 2003). Another reason for this binary focus is that for all firm pairs with a positive *HCF* value in our data, 88.8% have *HCF* equal to one. We note that because we have only a sample of employees for the firms in our data set, any observed positive *HCF* value in our data set may indicate a significant *HCF* between the two firms. Moreover, as we noted previously, the employees in our data set represent a more valued set of human capital than a random set of employees in the labor market. Hence, the migration of such employees would be of greater interest. For these reasons, a binary outcome variable for *HCF* with $\delta = 1$ is reasonable in itself. However, we also report results when $\delta = 2$ is used to dichotomize our outcome variable in Equation (4). This larger *HCF* threshold can indicate a stronger labor market competition relationship between two firms. In other words, the two different δ values constitute two different definitions of labor market competition. When $\delta = 1$, we identify all labor market competitors that have any *HCF* between them in a given year. When $\delta = 2$, we only identify the strong labor market competitors, and as seen in Table 5, this reduces the list of competitor pairs to a small fraction of the data. In addition, we also report, in Online Appendix E, prediction performances for a continuous interfirm competition variable, which is the same as the numeric *HCF* values between source-target pairs.

5.3. Predictive Models

Using the metrics constructed in the preceding section and as summarized in Table 4, we now describe the predictive models and corresponding results. We use observations from the years 2000 to 2012 for training the predictive models. We then evaluate the predictions from the different models for the observations in 2013 and 2014 (see Table 5 for a summary). Of the training data, we use observations in 2011 and 2012 as the validation set for hyperparameter tuning.

We include popular machine learning methods such as *K*-nearest neighbors (KNNs), regularized logistic regression, support vector machines (SVMs), and classification and regression tree (CART) as baseline models for prediction. Although linear and logistic regressions are popular for explanatory modeling

because of their transparent nature, the addition of a regularization component in these models allows them to be competitive on prediction performance with other more complex and opaque machine learning models. KNN, because of its simplistic implementation, also serves as a reasonable benchmark for the prediction problem. In addition, we experiment with predictive models that use an ensemble approach of training multiple models and then aggregate their output to lower the resulting prediction errors. These models include bootstrap aggregation or bagging of logistic regression (Bag(LR)) and bagging of support vector machines (Bag(SVMs)), as well as a tree-based ensemble method, random forest (Breiman 2001). Finally, we also include deep learning methods such as multi-layer perceptron (MLP) and convolutional neural network (CNN) as predictive models for the problem. We tune each of these methods by experimenting with various hyperparameter values using the validation data. The results of these tuning experiments and the consequent best-performing hyperparameters are presented in Online Appendix F. The results we present next are based on the best variation of each method identified through hyperparameter tuning and evaluated on held-out test data from the years 2013 and 2014.

5.4. Prediction Results

As shown in Table 4, we consider three types of metrics as predictors: economic, product overlap, and human capital overlap. For human capital overlap, we have proposed the skill-based labor overlap and the *HCF* network overlap, which are the focus of this study. However, the other economic and product overlap metrics serve as important control metrics so that we can tease out the predictive utility of our proposed human capital overlap metrics above and beyond that provided by the control metrics. With this goal, we evaluate our models first with just the economic metrics, followed by incrementally adding product overlap, labor overlap, and *HCF* network overlap to the predictor set.¹⁶

Table 6 shows the predictive performance of various machine learning methods (columns) with different sets of predictors (rows) in terms of the area under the receiver operating characteristics (ROC) curve (AUC). AUC is one of the most popular metrics for measuring the performance of classifiers. The AUC of a classifier is equivalent to the probability that a randomly chosen positive sample (competitors) will be ranked higher than a randomly chosen negative sample (noncompetitors), where the ranking is based on the predicted probabilities (Provost and Fawcett 2001, Fawcett 2004). For each model, we repeat the prediction experiment five times, where each time a random 80% of the training data is used to build a model. We evaluate the variation of a model's

Table 5. Data Set Summary

Data set	Size (firm pairs)	% (firm pairs with $Y = 1$)	
		$\delta = 1$	$\delta = 2$
Training (2000–2012)	65,176	23.1	2.0
Validation (2011–2012)	18,375	22.7	2.9
Test (2013–2014)	23,297	22.4	3.5

Table 6. Prediction Performance in Area Under the Receiver Operating Characteristics (ROC) Curve (AUC)

Feature set	KNN	LR	SVM	CART	Bag(LR)	Bag(SVM)	RF	MLP	CNN
Panel A: $\delta = 1$ (all competitors)									
Economic	0.655 (0.002)	0.572 (0.003)	0.596 (0.005)	0.617 (0.005)	0.570 (0.003)	0.583 (0.015)	0.692 (0.001)	0.572 (0.005)	0.629 (0.005)
Economic + product	0.658 (0.002)	0.587* (0.003)	0.609* (0.003)	0.619 (0.006)	0.585* (0.002)	0.568 (0.015)	0.701* (0.001)	0.585* (0.003)	0.651* (0.006)
Economic + product + labor	0.670* (0.003)	0.636* (0.001)	0.624* (0.004)	0.630* (0.004)	0.630* (0.001)	0.606* (0.015)	0.713* (0.001)	0.638* (0.004)	0.650 (0.005)
Economic + product + labor + network	0.803* (0.002)	0.777* (0.002)	0.829* (0.004)	0.845* (0.004)	0.746* (0.004)	0.814* (0.012)	0.890* (0.001)	0.782* (0.001)	0.770* (0.003)
Panel B: $\delta = 2$ (strong competitors)									
Economic	0.791 (0.004)	0.768 (0.002)	0.639 (0.019)	0.761 (0.009)	0.724 (0.001)	0.737 (0.008)	0.810 (0.002)	0.764 (0.001)	0.801 (0.003)
Economic + product	0.827* (0.003)	0.799* (0.003)	0.669 (0.009)	0.776 (0.010)	0.752* (0.001)	0.724 (0.015)	0.836* (0.002)	0.795* (0.002)	0.801 (0.003)
Economic + product + labor	0.858* (0.004)	0.851* (0.004)	0.724* (0.010)	0.820* (0.013)	0.840* (0.002)	0.742 (0.017)	0.861* (0.003)	0.848* (0.012)	0.801 (0.003)
Economic + product + labor + network	0.857 (0.003)	0.859 (0.003)	0.741* (0.010)	0.821 (0.010)	0.841* (0.001)	0.792* (0.019)	0.870* (0.002)	0.863* (0.001)	0.841* (0.004)

Notes. LR and SVM include a regularization term with $L2$ norm. Each algorithm is trained five times, where each time a random 80% of the training data are used for training. Standard errors of the five algorithms' performances on the test set are in parentheses. The best-performing models for $\delta = 1$ and 2 are highlighted in boldface. KNN, K-nearest neighbors; LR, logistic regression; SVM, support vector machine; CART, classification and regression tree; Bag(LR), bagging of logistic regressions; Bag(SVM), bagging of support vector machines; RF, random forest; MLP, multilayer perceptron; CNN, convolutional neural network.

* $p < 0.01$.

performances by calculating the standard error (included in parentheses) of its five performances on the test set. We test whether the performance improvement of a given model after incrementally adding features is statistically significant with the one-tailed, two-sample t -test. For example, from panel A of Table 6, we observe that the KNN's AUC with feature set (economic + product) is not statistically significantly better than its AUC with economic feature set alone. However, KNN with feature set (economic + product + labor) has a statistically significant improvement in AUC compared with its AUC with feature set (economic + product). The prediction experiment is performed twice, as shown in panels A and B of Table 6, with the δ in Equation (4) set to one (all identified labor market competitors) and two (strong labor market competitors), respectively. All the models are used for the classification problem where the outcome variable is binary, indicating a positive HCF value (of δ or less) or not. We would like to note that our main focus is on looking at one algorithm at a time, represented by different columns in Table 6. Given an algorithm, we want to understand the predictive performance improvements after adding our proposed metrics. In this way, we can understand the predictive utility of our proposed metrics across different state-of-the-art algorithms. The main observations from Table 6 are as follows:

- By comparing the predictive performances of models using the control metrics (i.e., economic + product overlap) with the models that additionally include the labor overlap metrics, we observe that models including the labor overlap metrics (*SkillTermSim* and *SkillTopicSim*) outperform models without them for $\delta = 1$ (panel A). Except for CNN, the differences in AUC values are all statistically significant (i.e., the addition of proposed labor overlap metrics is generally helpful for predicting all competitors across models). The predictive utility of skill-based labor overlap metrics is also strong for $\delta = 2$ (strong competitors). For almost all models, except CNN and Bag(SVM), we observe a sizable and statistically significant increase in AUC when the labor overlap metrics are included in addition to the control metrics (economic + product overlap). The results highlight the predictive utility of labor overlap metrics in their ability to capture all as well as just the strong labor market competitors.

- The predictive utility of HCF network overlap metrics is also clear from Table 6. In particular, we see large and statistically significant improvements in AUC when the HCF network overlap metrics are included for all models in panel A. Specifically, the improvement in performance can range between 18% (Bag(LR)) and 34% (Bag(SVM)) depending on the model. By contrast, when we consider only the strong labor market competitors, as shown in panel B,

the additional predictive power provided by the network overlap is not as large. In other words, when the labor competition between a pair of firms is strong, the similarity in explicit skills can provide most of the predictive utility beyond the control metrics. The tacit knowledge captured by the network overlap diminishes in terms of additional utility. This is in contrast to panel A ($\delta = 1$), which includes all labor market competitors, where tacit knowledge captured through network overlap provides greater additional utility. Strong labor market competitors are likely using similar technologies and hence require similar explicit knowledge inputs from labor. Hence, the predictive role of overlap in tacit skills (i.e., network overlap) diminishes for strong competitors after the similarity in explicit skills (i.e., labor overlap) is known.

- The best-performing models for $\delta = 1$ and 2 are highlighted in boldface in our result tables. Overall, the best-performing model is the ensemble-based random forest (RF) using all four types of predictors. This is consistent for both values of δ . In particular, for $\delta = 1$, it achieves an AUC of 0.89. We note that a random classifier would achieve an AUC of 0.5. Our best-performing model hence can significantly outperform a random baseline classifier. In particular, it is expected to identify a competitor pair over a noncompetitor pair without HCF correctly in the test set with a probability of 0.89. The RF with full feature

set is also the most effective in predicting strong labor market competitors, as shown in panel B of Table 6. Hence, by using carefully crafted human capital overlap metrics derived from employee skills and the labor supply network, along with RF models, we can provide strong predictive performance for identifying future labor market competitors.

Given the dynamic nature of firm labor market competition, we would like to separately evaluate the predictive performance of the models on just the new labor market competitors that are previously unknown. There is a set of firm pairs that exists only in our test set because their first HCF appears in 2013 or 2014. In other words, such firm pairs are not included in our training data, and they all have positive *HCF* values only in our test set. There are 3,062 and 150 such new firm pairs in the test set with their first observed *HCF* ≥ 1 and 2, respectively. To evaluate the predictive performances on these new firm pairs, we use the same set of models as used in Table 6. The output of these models is the probability of a firm pair being labor market competitors. We classify test firm pairs to be labor market competitors if the output probability from the model exceeds the prior probability (of labor market competitors) in the training data (0.231 for $\delta = 1$ and 0.020 for $\delta = 2$, as shown in Table 5). The proportion of new firm pairs that are correctly identified as labor market competitors is reported in Table 7.

Table 7. Proportion of New Future Competitors Identified

Feature set	KNN	LR	SVM	CART	Bag(LR)	Bag(SVM)	RF	MLP	CNN
Panel A: $\delta = 1$ (all competitors)									
Economic	0.699 (0.007)	0.441 (0.050)	0.424 (0.071)	0.341 (0.023)	0.410 (0.004)	0.459 (0.161)	0.759 (0.002)	0.392 (0.024)	0.576 (0.049)
Economic + product	0.609 (0.008)	0.361 (0.034)	0.307 (0.016)	0.312 (0.023)	0.343 (0.010)	0.364 (0.092)	0.758 (0.004)	0.337 (0.015)	0.655 (0.029)
Economic + product + labor	0.571 (0.007)	0.350 (0.022)	0.257 (0.022)	0.312 (0.010)	0.341 (0.014)	0.399 (0.112)	0.734 (0.005)	0.352 (0.018)	0.788* (0.039)
Economic + product + labor + network	0.924* (0.005)	0.765* (0.018)	0.888* (0.003)	0.838* (0.009)	0.814* (0.030)	0.885* (0.024)	0.964* (0.001)	0.934* (0.009)	0.787 (0.057)
Panel B: $\delta = 2$ (strong competitors)									
Economic	0.405 (0.048)	0.346 (0.164)	0.636 (0.037)	0.168 (0.047)	0.028 (0.001)	0.702 (0.116)	0.517 (0.009)	0.113 (0.010)	0.128 (0.065)
Economic + product	0.407 (0.007)	0.390 (0.078)	0.602 (0.026)	0.253 (0.081)	0.210* (0.027)	0.727 (0.113)	0.516 (0.009)	0.113 (0.006)	0.128 (0.063)
Economic + product + labor	0.375 (0.029)	0.455* (0.058)	0.595 (0.031)	0.264 (0.041)	0.357* (0.016)	0.683 (0.032)	0.509 (0.015)	0.102 (0.026)	0.125 (0.061)
Economic + product + labor + network	0.499* (0.017)	0.493 (0.059)	0.645 (0.069)	0.343 (0.095)	0.352 (0.018)	0.817* (0.046)	0.622* (0.020)	0.126 (0.009)	0.142 (0.045)

Notes. LR and SVM include a regularization term with *L2* norm. Each algorithm is trained five times, where each time a random 80% of the training data are used for training. Standard errors of the five algorithms' performances on the test set are in parentheses. The best-performing models for $\delta = 1$ and 2 are highlighted in boldface. KNN, K-nearest neighbors; LR, logistic regression; SVM, support vector machine; CART, classification and regression tree; Bag(LR), bagging of logistic regressions; Bag(SVM), bagging of support vector machines; RF, random forest; MLP, multilayer perceptron; CNN, convolutional neural network.

* $p < 0.01$.

From Table 7, we again observe the greatly improved performance of predicting new labor market competitors when human capital overlap metrics are included in addition to the control metrics. Specifically, the network overlap metrics provide strong predictive utility in identifying new firm competitor pairs for both $\delta = 1$ and 2. For example, when weaker competitors are included (i.e., $\delta = 1$), we see a 200%–300% better performance for several machine learning models that use of the network overlap metrics on top of control metrics and labor overlap metrics. When all four types of metrics are used (economic + product + labor + network), the RF model is able to correctly identify 96.4% of the new competitors. Overall, as one might expect, it is more challenging to predict new strong competitors (i.e., $\delta = 2$). Nevertheless, Bag(SVM) is able to capture, on average, 81.7% of the new strong competitors. Hence, the evidence clearly indicates the utility of the proposed human capital metrics for predicting labor market competitors, including previously unseen ones. We have replicated the analysis using another set of data collected in 2016, and the main results in terms of the predictive utility of our proposed metrics remain the same (see Online Appendix G).

6. Discussion

6.1. 2D Competitor Analysis

We now illustrate how our proposed labor market overlap metrics and resulting predictive models can be incorporated into the 2D competitor analysis (Chen 1996, Peteraf and Bergen 2003, Markman et al. 2009), as shown in Figure 1, and support business intelligence-gathering efforts. To measure a firm pair's product market overlap, we take the average of their *Busdesc TermSim* and *BusdescTopicSim*. The two measures provide somewhat different information because they focus on different levels of granularity of skills. Hence, the average helps to combine the two pieces of information on interfirm product overlap. Similarly, we take the average of a firm pair's *SkillTermSim* and *SkillTopicSim* to measure their labor overlap. These skill-based labor overlap metrics are a direct measure of the similarity in the explicit knowledge endowment of two firms and are consistent with the measure on the y -axis in terms of firms' similarity in the product market that is based on the textual descriptions of firms. In addition, in this section, only the predictions based on the RF model with the full feature set are considered for analysis.

For ease of visualization, Figure 7 plots only the strong competitors (i.e., $\delta = 2$), and the size of the dot indicates the size of the HCF between them. The dashed lines plot the median values of the product overlap and the labor overlap given all firm pairs in the test set. Hence, the median lines divide the plot into four quadrants, as initially conceptualized in

Figure 7. The 2D Competition Plot for All Test Firm Pairs

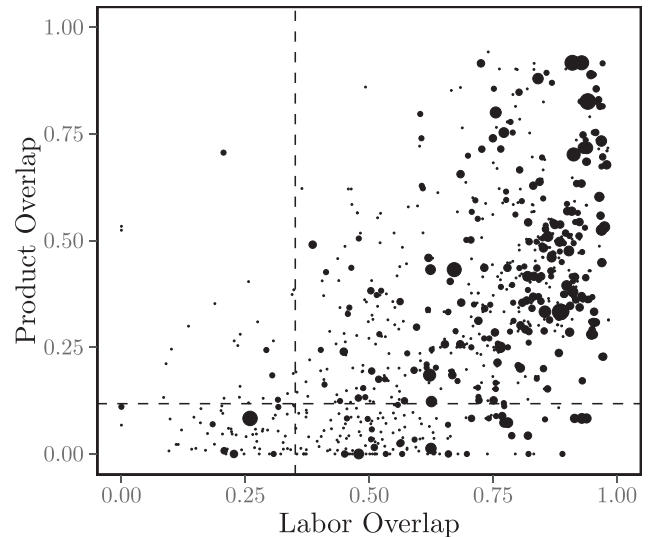


Figure 1. We find that most of the labor market competition appears in the upper-right quadrant, which corresponds to direct competitors that have a large overlap in both product and labor dimensions.

When the RF model's prediction performances ($\delta = 2$) are considered separately for the four quadrants, the model's AUC is 0.839 for the upper-right quadrant (direct competitors), 0.775 for the upper-left quadrant (indirect competitors), 0.718 for the lower-left quadrant (weak competitors), and 0.765 for the lower-right quadrant (potential competitors). The best performances are achieved among firm pairs that are direct competitors and hence present various cues to their competition through product and human capital overlap. In contrast, the prediction performance is lowest (although still reasonable) for weak competitors, where firm pairs are expected to have smaller overlap in both outputs (products) and inputs (human capital). The varying levels of AUC for different quadrants inform users about the confidence they can have on the predictions based on where the unseen firm pairs of interest (to the user) fall.

The visualization using the product and labor overlap metrics, as shown in Figure 7, also allows users to identify interesting exceptions. For example, there is a large $HCF = 17$ from Hewlett-Packard (HP) to General Motor (GM) in 2013, which is the large dot in the lower-left quadrant (weak competitors) of Figure 7. This is a firm pair with a low product overlap of 0.08 and a low labor overlap of 0.26. Our overall best-performing model, RF, correctly predicts the firm pair as labor market competitors ($\delta = 2$) using prior probability from training data as a threshold. This observed large HCF probably reflects GM's agreement with HP to hire up to 3,000 HP employees already working on GM's business starting in 2012.¹⁷ Although not the focus of this work, it may reflect GM's

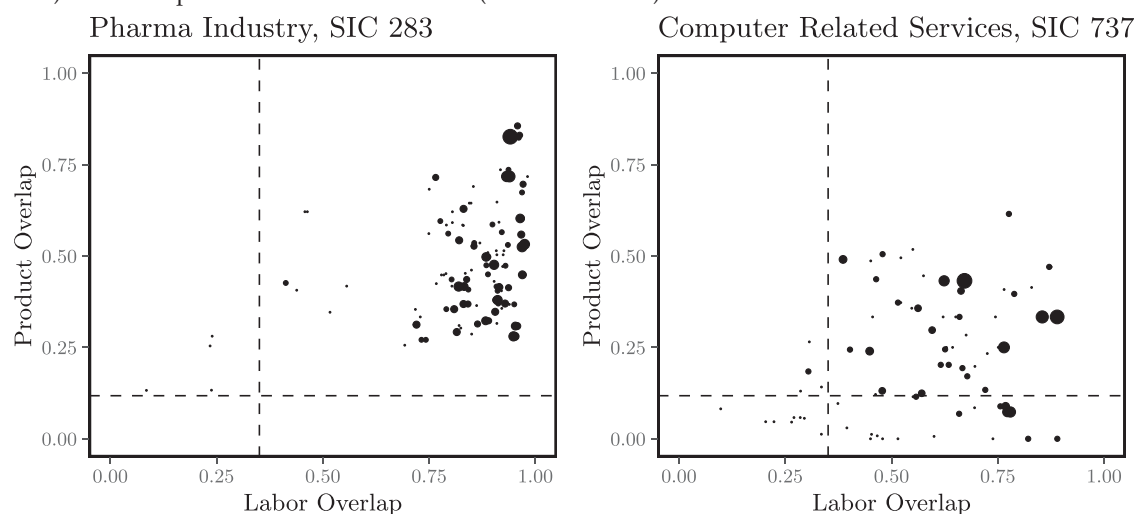
strategy to accelerate software innovation through insourcing IT talent at a time when cars were becoming increasingly computerized.¹⁸ For HP, GM is a weak competitor, as shown in Figure 7. However, GM is not irrelevant to HP because it is a client of HP. As a result, the departure of employees to GM can be beneficial to HP: the movement of employees can facilitate the creation and strengthening of further business relationships between HP and GM, and the moving of employees' knowledge and work practices can make future interorganizational endeavors more efficient (Somaya et al. 2008). Hence, for HP, the implications and strategic responses will be different between losing employees to a weak competitor (GM) versus a direct competitor (Apple or Dell). This example shows how the metrics and prediction framework proposed in this paper can enrich firm competitor analysis and provide insights that may be largely ignored if competitor analysis is performed in a context devoid of the four quadrants in Figure 1. Depending on the quadrant where a competitor pair is predicted or identified, a firm may have vastly different strategies for acting on the discovery.

The 2D competition analysis can be performed for each industry separately, providing a more nuanced picture of the industry-specific competitor landscape. Figure 8 shows such analysis. Specifically, the left panel in Figure 8 shows a subset of firm pairs where the source firm is a drug (pharmaceutical) firm (SIC Code starting in 283). The RF model can achieve an AUC of 0.915 for such firm pairs in our test set. Clearly, the labor market competitors faced by pharmaceutical firms are mostly other firms with high product and labor market overlap simultaneously because most firm pairs are in the upper-right quadrant of the subplot. In fact, 93.1% of the firm pairs in the left panel of Figure 8 include both the source and target firms

from the drug industry (SIC Code starting with 283). The few competitor firm pairs in the upper-left quadrant are with target firms such as ExxonMobil (oil industry), Mondelez (food industry), and Altria Group (tobacco industry). In addition, the left panel of Figure 8 shows no firm pairs from a pharmaceutical firm to a firm with low product market overlap. This may be because of the fact the knowledge required by the pharmaceutical firms is very much industry specific; hence, our observation here can also be described as an industry-level *boundary* in the mobility pattern of their employees (Farjoun 1994).

The right panel of Figure 8 shows a different competitive landscape for the source firms that are in the industry of computer-related services (SIC Code starting in 737; e.g., Google, IBM, and Microsoft). The RF model can achieve an AUC of 0.882 for such firm pairs in our test set. As expected, firm pairs with both high product and labor market overlap also have higher levels of HCF. However, the labor market competitors occupy a different subspace of the 2D competition plot than those from the pharmaceutical industry. Specifically, compared with the pharmaceutical industry (left subplot), only 39.0% of the firm pairs in computer-related services include both the source and target firms from the same industry (SIC Code starting with 737). In particular, the large HCF dots in the lower-right quadrant are from firms in this industry to Amazon. Amazon, despite its substantial cloud business, is a large consumer product retailer in terms of its products or output. Hence, it has a low product overlap with many firms in computer-related services. However, it has a business model with IT at its core and hence has human capital needs that are similar to those of firms in computer-related services. In general, we observe HCF from the computer-related services industry to a wide variety of other

Figure 8. The 2D Competition Analysis for Test Firm Pairs: (Left Panel) with Drug Firms (SIC Code = 283) As Source Firms and (Right Panel) with Computer-Related Service Firms (SIC Code = 737) As Source Firms



industries (e.g., Visa, UnitedHealth Group, Lowes, and General Dynamics). The diverse target firms for the HCF from firms in the computer-related services industry are a reflection of the more general applicability of computer-related knowledge than the pharmaceutical-related knowledge to other industries (Joseph et al. 2012). Taken together, the industry-level analysis presents an example of how a competitor analysis incorporating both product and labor market dimensions can provide a broader view of human capital development in different industries. The varying levels of AUC across different industries indicate the varying levels of confidence that the user can have on the predictive model given an industry. Although the predictive performance is strong for both pharmaceutical and computer-related services industries, the user can have even greater confidence in the prediction over unseen data when using the model in the pharmaceutical industry.

6.2. Managerial Implications

An important managerial implication of our work is in identifying and improving awareness of a firm's competitors for human capital (Chen et al. 2007). Competitor identification and awareness are important because they are necessary precursors to competitor analysis and strategy (i.e., all the subsequent analysis stands on the foundation of competitor identification; Pant and Sheng 2015). Previous literature has well documented that managers may have "myopia" or "blind spots" when it comes to identifying a firm's product market competitors (Zajac and Bazerman 1991, Walker et al. 2005, Pant and Sheng 2015).¹⁹ We can expect such myopia to be worse for identifying a firm's competitors in the labor market. This is because although plenty of product-level data about firms are available, there are relatively few data sources for managers to observe other firms' labor market behaviors. Because our labor market competitor prediction includes a large number of firms across industries, it can help improve managers' awareness of their firm's labor market competitors, especially from distant product markets. Also, after our methodology identifies a pair of firms as future labor market competitors, we can apply the 2D competitor analysis framework, which can help managers make strategic decisions beyond HR, such as product development and customer relationship management (Peteraf and Bergen 2003, Somaya and Williamson 2008, Markman et al. 2009). We note that such a 2D competitor analysis is possible as a result of our proposed skill-based labor overlap metrics that serve as the measurement for the horizontal axis. Moreover, our proposed HCF prediction and 2D competitor analysis can provide a global view of employee mobility across firms. Existing studies have explored the impact of external firms' employee migrations on a focal firm's

business (Carnahan and Somaya 2015), for example, how a supplier firm is affected when a buyer hires employees from the supplier's competitors (Carnahan and Somaya 2013) and how the knowledge of anticipated employee mobility affects firm acquisition value and likelihood (Younge et al. 2015). Such externalities of the firm's hiring decisions on other firms further highlight the need for a global view of employee migrations. Hence, the implications of predicting future employee migrations go beyond the managers at the focal firms themselves. In Online Appendix A, we illustrate further a single-firm-use case of a manager or analyst visualizing the predicted labor market competitors in the 2D space. It is clear that the predicted probabilities, along with their 2D context concerning product and labor overlap, provide a useful mechanism for users to hone in on future labor market competitors and their strategic implications.

6.3. Academic Implications

In addition to the managerial implications discussed, this study also contributes to the academic literature in the following manner. We note that there exist theoretical/qualitative studies regarding how, why, and the conditions under which firms compete for human capital and its relationship with firm product market competition (Chen 1996, Peteraf and Bergen 2003, Markman et al. 2009). However, because of constraints on data availability and the associated lack of popularity of many data science methods, there have been few empirical or predictive analytics studies on firm labor market competition, especially for a large number of firms. Hence, in this study, we propose new metrics of firm human capital endowment and overlap and validate our proposed metrics by evaluating their effectiveness in predicting the firm's future labor market competitors. Our proposed human capital endowment distributions (e.g., skill vectors, upstream and downstream vectors) and corresponding overlap measures provide a rich set of quantitative representations to test implications of the resource-based view of firms. Also, with the proposed human capital overlap metrics and the prediction of HCF, we can operationalize the 2D competition analysis described by previous qualitative studies with significant potential for practical utility (Markman et al. 2009). The operationalization of the 2D competition analysis makes possible future empirical studies that use the relative positioning of firm pairs in different quadrants. For example, it will be interesting to study the transition of firm pair relationships across quadrants of Figure 1 over time. Such investigations can advance our understanding of firm competitive dynamics with a multimarket perspective that goes beyond existing qualitative notions (Markman et al. 2009).

7. Conclusion

In this study, we focus on the prediction of future labor market competition. However, it is important to note the paucity of empirical work even on the contemporaneous identification of labor market competitors in previous literature. Hence, this study is not just the first in terms of addressing the problem of predicting future labor market competitors, but it is also a first in terms of suggesting metrics of human capital overlap that can help in even contemporaneous identification of labor market competition. We are able to operationalize the notions of (human) resource bundles, as suggested by RBV, by viewing them as distributions over explicit and tacit knowledge. We contend that such operationalization of firm-level human capital and consequent overlap metrics have implications for the general area of strategic human capital and HR management literature beyond the prediction problem addressed here.

We use the public profiles of more than 89,000 employees and construct various metrics for the human capital overlap between firms. The granular individual-level skill terms allow us to create skill-based labor overlap measures that indicate the interfirm similarity in their explicit knowledge base. Also, because the HCF between firms reflects the flow of both explicit and tacit knowledge, the resulting network structure is leveraged to measure more global metrics of human capital overlap. We find that our proposed labor- (skill) and network-based human capital overlap metrics are critical to good predictive performance for the prediction of future labor market competitors. By proposing human capital overlap metrics and using them for predicting future labor market competitors, our paper fulfills the long-existing need for a comprehensive firm competitor analysis with both the product and labor dimensions of interfirm overlap (Chen 1996, Peteraf and Bergen 2003). We also discuss the nuanced analysis that the two dimensions allow while focusing on labor market competition.

Applying data analytics for talent acquisition and retention has been identified as one of the most urgent challenges facing HR leaders around the world. However, it is also one of the challenges that firms are least ready for (Deloitte 2015). In addition to the firm-level competitor analysis, the HCF prediction task in our study can help narrow such a capability gap between the urgency and readiness of data-driven HR management. For example, our prediction framework can be used to form the basis of a targeted recruitment strategy. On the one hand, for an HR manager, the predictive models can provide a list of firms as targets for future hiring so that the rate of successful hires can be increased and the cost of hiring reduced. On the other hand, our predictive models can also identify a

set of firms that may target a particular firm for hiring its employees. Such information can be of strategic value for a firm for at least two reasons. First, it can help HR design a more effective talent retention program. Second, tracking where employees likely may be leaving for can be crucial for a company's future strategic development (Somaya and Williamson 2008). Moreover, the knowledge of which quadrant (see Figure 1) the future labor market competitors fall into allows a manager or analyst to apply different strategies for different labor market competitors. Finally, we note that our methodology uses only publicly available data and hence can be implemented with relative ease.

We seed our data with employees of S&P 100 firms in 2015 and then extend the analysis to encompass other firms (3,467 firms in total) in which they were previously employed. Because search engines carefully guard the details of their ranking algorithms, we do not exactly know how Yahoo BOSS identifies its top 1,000 search results. We do not expect the ranking to be random. Instead, top results are likely to have a bias toward more popular and potentially higher-valued employees within a firm. In this manner, as we noted earlier, our analysis can be seen as focusing on the movement of more-valued human capital than employees of a random set of firms. However, the movement of such higher-valued human capital would be of considerable interest in assessing labor market competition between firms. In the future, the analysis can be extended by seeding it with the employees of an even larger set of firms so as to have a greater diversity in terms of the quality of human capital. Another future direction is to predict labor market competition at a more granular skill level. For example, in Online Appendix D, we show how our proposed metrics can be used to predict interfirm HCF for specific skill categories discovered by the LDA (Section 3.1). The skill-specific prediction can provide HR managers with information to provide retention incentives to employees in an even more targeted manner. Also, our proposed human capital overlap metrics may be used for future product market competitor prediction. As a result of the relationship between firms' product market and labor market interactions (Markman et al. 2009, Younge et al. 2015), it is valuable to identify long-run competitors that may manifest many years after talent acquisition (e.g., despite hiring from Tesla, Apple has yet to produce a competing car). Such a long-run prediction (although being challenging at various levels such as data acquisition) will have a high impact. Finally, the network analysis presented here is at the level of firms, but it can be applied to a context where the nodes are industries (at various levels of granularity) or states (as well as nations)

instead of firms. Hence, the proposed framework can be extended to predicting interindustry, interstate, or even international competition for human capital. In summary, web footprints of employees are providing rich information on labor market competition that can provide predictive utility and insights across the economy.

Acknowledgments

The authors thank the editors and the anonymous reviewers for their helpful comments and suggestions. They are also thankful for the conversations with various participants at seminars, conferences, and workshops.

Endnotes

¹ See <https://www.cnet.com/news/amazon-com-wal-mart-settle-lawsuit/>.

² See <https://money.cnn.com/2018/05/15/news/companies/tech-banking-amazon-apple-facebook/index.html>.

³ Note that Microsoft appears among both the upstream and downstream firms for Amazon. This is expected, and hence the labor market is viewed as a network and not a linear chain.

⁴ See <https://www.linkedin.com/pulse/linkedin-industry-rankings-see-which-tops-list-joshua-waldman>.

⁵ The keywords in our search query include “background,” “experience,” “education,” “skills,” and “current: <firm name>.” The first four keywords represent the key Hypertext Markup Language content in a profile page that helps us identify profile pages. The “current: <firm name>” keyword is used to identify employees in one of the S&P 100 firms. Yahoo BOSS returns a maximum of 1,000 results for a search query.

⁶ Securities and Exchange Commission filings of companies are the primary source of Compustat’s data. The Compustat database includes only publicly held firms, both active and inactive. We only use the primary SIC Code of firms.

⁷ Such filtering of terms is common in information retrieval to avoid misspellings.

⁸ The predictive performances are similar with other similarity metrics, such as Jaccard’s similarity. We show results with other operationalizations of our proposed metrics in Online Appendix E.

⁹ We find similar prediction performances with a much larger number of topics (Online Appendix E).

¹⁰ See http://www.osha.gov/pls/imis/sic_manual.html.

¹¹ We would like to highlight the previous use of SIC Codes in a more rigid framework where two firms are considered similar if they have the same SIC Code (or some arbitrarily defined set of digits within it, e.g., the first two digits) and dissimilar otherwise. In contrast, we allow SIC Codes to be similar to various degrees based on their well-defined hierarchy.

¹² We note that firms can sometimes acquire new skills or knowledge from unrelated firms. Our metric does not exclude such cases because looking at the network similarity only provides an “on average” cue to human capital overlap and does not assume future hiring from related or unrelated firms.

¹³ We find similar prediction results with another network community detection algorithm (Online Appendix E).

¹⁴ We identify undergraduate degrees, such as “BA,” “BS,” or a degree with the word “Bachelor.” If no such degree is identified, the last degree is used; otherwise, this employee is excluded from further calculation of this variable.

¹⁵ We use the Quacquarelli Symonds (QS) university rank in 2015 as the ranking of universities (<http://www.topuniversities.com/qs-world-university-rankings>).

¹⁶ We show effects of individual and subgroup features’ prediction performances in Online Appendix E.

¹⁷ See https://media.gm.com/media/us/en/gm/home.detail.html/content/Pages/news/us/en/2012/Oct/1018_it_transformation.html.

¹⁸ See <https://www.technologyreview.com/s/506746/with-computerized-cars-ahead-gm-puts-it-outsourcing-in-the-rearview-mirror/>.

¹⁹ Explanations for such managerial myopia include limitations in firm resources, bounded rationality, and cognitive biases of managers (Peteraf and Bergen 2003).

References

- Almeida P, Kogut B (1999) Localization of knowledge and the mobility of engineers in regional networks. *Management Sci.* 45(7): 905–917.
- Ambrosini V, Bowman C (2001) Tacit knowledge: Some suggestions for operationalization. *J. Management Stud.* 38(6):811–829.
- Arthur MB (2014) The boundaryless career at 20: Where do we stand, and where can we go? *Career Development Internat.* 19(6):627–640.
- Axelrod EL, Handfield-Jones H, Welsh TA (2001) War for talent, part two. *McKinsey Quart.* 2(2):9–12.
- Barney J (1991) Firm resources and sustained competitive advantage. *J. Management* 17(1):99–120.
- Bartlett CA, Ghoshal S (2002) Building competitive advantage through people. *MIT Sloan Management Rev.* 43(2):34.
- Beckert J (2010) Institutional isomorphism revisited: Convergence and divergence in institutional change. *Sociol. Theory* 28(2): 150–166.
- Beechler S, Woodward IC (2009) The global “war for talent.” *J. Internat. Management* 15(3):273–285.
- Bergen M, Peteraf MA (2002) Competitor identification and competitor analysis: A broad-based managerial approach. *Managerial Decision Econom.* 23(4–5):157–169.
- Blei DM (2012) Probabilistic topic models. *Comm. ACM* 55(4):77–84.
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3(January):993–1022.
- Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J. Statist. Mech. Theory Experiment* 2008(10):P10008.
- Breiman L (2001) Random forests. *Machine Learn.* 45(1):5–32.
- Cappelli P (2008) Talent management for the twenty-first century. *Harvard Bus. Rev.* 86(3):74.
- Carnahan S, Somaya D (2013) Alumni effects and relational advantage: The impact on outsourcing when a buyer hires employees from a supplier’s competitors. *Acad. Management J.* 56(6):1578–1600.
- Carnahan S, Somaya D (2015) The other talent war: Competing through alumni. *MIT Sloan Management Rev.* 56(3):14.
- Chambers EG, Foulon M, Handfield-Jones H, Hankin SM, Michaels EG, et al. (1998) The war for talent. *McKinsey Quart.* 3:44–57.
- Chen M-J (1996) Competitor analysis and interfirm rivalry: Toward a theoretical integration. *Acad. Management Rev.* 21(1):100–134.
- Chen M-J, Miller D (2012) Competitive dynamics: Themes, trends, and a prospective research platform. *Acad. Management Ann.* 6(1):135–210.
- Chen M-J, Su K-H, Tsai W (2007) Competitive tension: The awareness-motivation-capability perspective. *Acad. Management J.* 50(1): 101–118.
- Cliffe S (1998) Winning the war for talent. *Harvard Bus. Rev.* 76(5): 18–20.
- Davenport TH, Harris J, Shapiro J (2010) Competing on talent analytics. *Harvard Bus. Rev.* 88(10):52–58.

- Davison BD (2000) Topical locality in the web. *Proc. 23rd ACM SIGIR Conf.* (ACM, New York), 272–279.
- Deloitte (2015) Global human capital trends 2015. Technical report. Deloitte, https://www2.deloitte.com/content/dam/Deloitte/na/Documents/human-capital/na_DUP_GlobalHumanCapitalTrends2015.pdf.
- DiMaggio PJ, Powell WW (1983) The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *Amer. Sociol. Rev.* 48(2):147–160.
- Doidge C, Karolyi GA, Stulz RM (2017) The US listing gap. *J. Financial Econom.* 123(3):464–487.
- Farjoun M (1994) Beyond industry boundaries: Human expertise, diversification and resource-related industry groups. *Organ. Sci.* 5(2):185–199.
- Fawcett T (2004) ROC graphs: Notes and practical considerations for researchers. *Machine Learn.* 31(1):1–38.
- Gardner TM (2002) In the trenches at the talent wars: Competitive interaction for scarce human resources. *Human Resource Management* 41(2):225–237.
- Gardner TM (2005) Interfirm competition for human resources: Evidence from the software industry. *Acad. Management J.* 48(2): 237–256.
- Grant RM (1996a) Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organ. Sci.* 7(4):375–387.
- Grant RM (1996b) Toward a knowledge-based theory of the firm. *Strategic Management J.* 17(S2):109–122.
- Guerrero OA, Axtell RL (2013) Employment growth through labor flow networks. *PLoS One* 8(5):e60808.
- Hatch NW, Dyer JH (2004) Human capital and learning as a source of sustainable competitive advantage. *Strategic Management J.* 25(12):1155–1178.
- Higgins T, Hull D (2015) Want Elon Musk to hire you at Tesla? Work for Apple. *Bloomberg Bus.* (February 5), <https://www.bloomberg.com/news/articles/2015-02-05/want-elon-musk-to-hire-you-at-tesla-work-for-apple>.
- Hitt MA, Bierman L, Shimizu K, Kochhar R (2001) Direct and moderating effects of human capital on strategy and performance in professional service firms: A resource-based perspective. *Acad. Management J.* 44(1):13–28.
- Hom PW, Lee TW, Shaw JD, Hausknecht JP (2017) One hundred years of employee turnover theory and research. *J. Appl. Psych.* 102(3):530–545.
- Hoopes DG, Madsen TL, Walker G (2003) Guest editors' introduction to the special issue: Why is there a resource-based view? Toward a theory of competitive heterogeneity. *Strategic Management J.* 24(10):889–902.
- Horton JJ, Tambe P (2015) Labor economists get their microscope: Big data and labor market analysis. *Big Data* 3(3):130–137.
- Joseph D, Boh WF, Ang S, Slaughter SA (2012) The career paths less (or more) traveled: A sequence analysis of it career histories, mobility patterns, and career success. *MIS Quart.* 36(2):427–452.
- Kuhn PJ (2014) The internet as a labor market matchmaker. IZA World of Labor. Accessed October 30, 2020, <https://wol.iza.org/uploads/articles/18/pdfs/internet-as-a-labor-market-matchmaker.pdf>.
- Ma Z, Pant G, Sheng ORL (2011) Mining competitor relationships from online news: A network-based approach. *Electronic Commerce Res. Appl.* 10(4):418–427.
- Manning C, Raghavan P, Schütze H (2010) Introduction to information retrieval. *Natl. Language Engrg.* 16(1):100–103.
- Markman GD, Gianiodis PT, Buchholtz AK (2009) Factor-market rivalry. *Acad. Management Rev.* 34(3):423–441.
- McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Rev. Sociol.* 27(2001): 415–444.
- Newbert SL (2007) Empirical research on the resource-based view of the firm: An assessment and suggestions for future research. *Strategic Management J.* 28(2):121–146.
- Nyberg AJ, Moliterno TP, Hale D Jr, Lepak DP (2014) Resource-based perspectives on unit-level human capital: A review and integration. *J. Management* 40(1):316–346.
- Pant G, Sheng ORL (2015) Web footprints of firms: Using online isomorphism for competitor identification. *Inform. Systems Res.* 26(1):188–209.
- Pant G, Srinivasan P (2013) Status locality on the web: Implications for building focused collections. *Inform. Systems Res.* 24(3):802–821.
- Peteraf MA, Bergen ME (2003) Scanning dynamic competitive landscapes: A market-based and resource-based framework. *Strategic Management J.* 24(10):1027–1041.
- Ployhart RE, Weekley JA, Ramsey J (2009) The consequences of human resource stocks and flows: A longitudinal examination of unit service orientation and unit effectiveness. *Acad. Management J.* 52(5):996–1015.
- Provost F, Fawcett T (2001) Robust classification for imprecise environments. *Machine Learn.* 42(3):203–231.
- Rosenkopf L, Almeida P (2003) Overcoming local search through alliances and mobility. *Management Sci.* 49(6):751–766.
- Schilling MA, Phelps CC (2007) Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Sci.* 53(7):1113–1126.
- Schmutte IM (2014) Free to move? A network analytic approach for learning the limits to job mobility. *Labour Econom.* 29(2014): 49–61.
- Shi Z, Lee GM, Whinston AB (2016) Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Quart.* 40(4):1035–1056.
- Shmueli G, Koppius OR (2011) Predictive analytics in information systems research. *MIS Quart.* 35(4):553–572.
- Somaya D, Williamson IO (2008) Rethinking the “war for talent.” *MIT Sloan Management Rev.* 49(4):29–34.
- Somaya D, Williamson IO, Lorinkova N (2008) Gone but not lost: The different performance impacts of employee mobility between cooperators vs. competitors. *Acad. Management J.* 51(5): 936–953.
- Song J, Almeida P, Wu G (2003) Learning-by-hiring: When is mobility more likely to facilitate interfirm knowledge transfer? *Management Sci.* 49(4):351–365.
- Takeuchi R, Lepak DP, Wang H, Takeuchi K (2007) An empirical examination of the mechanisms mediating between high-performance work systems and the performance of Japanese organizations. *J. Appl. Psych.* 92(4):1069.
- Tambe P, Hitt LM (2011) The productivity of information technology investments: New evidence from IT labor data. *Inform. Systems Res.* 23(3 pt 1):599–617.
- Walker BA, Kapelani D, Hutt MD (2005) Competitive cognition. *MIT Sloan Management Rev.* (July 15), <https://sloanreview.mit.edu/article/competitive-cognition/>.
- Weiner C (2005) The impact of industry classification schemes on financial research. Preprint, submitted December 20, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=871173.
- Wigglesworth R (2015) Hedge funds poach computer scientists from Silicon Valley. *Financial Times* (November 22), <https://www.ft.com/content/856bd92c-8fb0-11e5-8be4-3506bf20cc2b>.
- Witten IH, Frank E (2005) *Data Mining: Practical Machine Learn. Tools and Techniques*, 2nd ed. (Morgan Kaufmann Publishers, San Francisco).
- Wright PM, McMahan GC (2011) Exploring human capital: Putting ‘human’ back into strategic human resource management. *Human Resource Management J.* 21(2):93–104.
- Wright PM, Coff R, Moliterno TP (2014) Strategic human capital: Crossing the great divide. *J. Management* 40(2):353–370.

- Wright PM, McMahan GC, McWilliams A (1994) Human resources and sustained competitive advantage: A resource-based perspective. *Internat. J. Human Resource Management* 5(2):301–326.
- Wu L, Jin F, Hitt LM (2018) Are all spillovers created equal? a network perspective on information technology labor movements. *Management Sci.* 64(7):3168–3186.
- Younge KA, Tong TW, Fleming L (2015) How anticipated employee mobility affects acquisition likelihood: Evidence from a natural experiment. *Strategic Management J.* 36(5):686–708.
- Zajac EJ, Bazerman MH (1991) Blind spots in industry and competitor analysis: Implications of interfirm (mis)perceptions for strategic decisions. *Acad. Management Rev.* 16(1):37–56.